IMPROVING DISTRIBUTION SYSTEM MODEL ACCURACY BY LEVERAGING UBIQUITOUS SENSORS

A Dissertation
Presented to
The Academic Faculty

by

Jouni Peppanen

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology
May 2016

IMPROVING DISTRIBUTION SYSTEM MODEL ACCURACY BY LEVERAGING UBIQUITOUS SENSORS

Approved by:

Dr. Santiago Grijalva, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Ronald G. Harley, Chair
School of School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Maryam Saeedifard
School of School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. A. P. Meliopoulos
School of School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Shabbir Ahmed
School of Industrial and Systems Engineering
Georgia Institute of Technology

Date Approved: March 29, 2016

To Anton, Jan, and Natallia

"We do not inherit the Earth from our ancestors,
we borrow it from our children"

ACKNOWLEDGEMENTS

I would like to thank my PhD research advisor Dr. Santiago Grijalva for his guidance and support throughout my PhD research. From him, I have learned a great deal about conducting high-quality research, effectively communicating ideas to others, and pursuing your vision.

I would also like to thank the rest of my committee members Dr. Ronald G. Harley, Dr. Maryam Saeedifard, Dr. A. P. Meliopoulos, and Dr. Shabbir Ahmed for their valuable feedback and time in improving the quality of this dissertation.

During my PhD research, I have been very fortunate to be part of a great team. I would like to thank all my colleagues at the ACES Laboratory for providing a stimulating and fun work environment. My special thanks go to Matthew J. Reno, John Seuss, Xiaochen Zhang, Jose Grimaldo, and Kyle Coogan. Above others, I would like to show my gratitude to Matthew J. Reno who has been closely involved in developing many of the ideas presented in this dissertation. I would also thank Robert Broderick from the Sandia National Laboratories for his valuable feedback on the distribution system secondary circuit parameter and topology estimation work.

I would also like to express my gratitude to Fortum Foundation, Finnish Cultural Foundation, Fulbright Commission of Finland, Finnish Technology Advancement Foundation, and Georgia Tech for financially supporting my PhD studies and research.

I am sincerely grateful to my parents Veikko and Hilkka Peppanen; my brother Toivo Peppanen; my uncle Olli Peppanen; and many other members of my extended family for their encouragement and support. My deepest gratitude belongs to my wife Natallia for her love, encouragement, support, and understanding. I am thankful for you and our boys Anton and Jan for putting things into perspective.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

NOMENCLATURE

AC	Alternating current
ACES	Advanced Computational Electricity Systems Laboratory at Georgia Tech
ADMS	Advanced distribution management systems
AHU	Aho, Hopcroft, and Ullman algorithm
AMI	Advanced metering infrastructure
AMR	Advanced meter reading
ANSI	American National Standards Institute
BP	Best practice data imputation method
CLS	Linearly constrained least squared estimator
CT	Current transformer
CVR	Conservation voltage reduction
DER	Distributed energy resource
DG	Distributed generation
DOE	Department of Energy
DOY	Day of the year
DIN	Data imputation and nowcasting
DR	Demand response
DSPE	Distribution system secondary circuit parameter estimation
DSSE	Distribution system state estimation
DSTE	Distribution system secondary circuit topology estimation
EIT	Electrically identical topology
EMS	Energy management system
EPR	Ethylene propylene rubber

EV	Electric vehicle
FLISR	Fault location, isolation and service restoration
GB	Gigabyte
GIS	Geographical information system
GW	Gigawatt (AC)
HA	Historic average data imputation method
HH	Hour of the day
HV	High-voltage
IEA	International Energy Agency
kW	kilowatt (AC)
kVAr	kilovolt-ampere reactive (AC)
kVA	kilovolt-ampere (AC)
LAV	Least absolute value estimator
LI	Linear interpolation
LRPE	Linear regression parameter estimation algorithm
LRTE	Linear regression joint topology and parameter estimation algorithm
LTC	On-load tap changer
LV	Low-voltage
MAE	Mean absolute error
MAPE	Mean absolute percentage error
MBE	Mean bias error
MDMS	Meter data management system
MLE	Maximum likelihood estimation
MM	Minute of the hour
MSE	Mean squared error
MV	Medium-voltage

MW	Megawatt
OLS	Ordinary least squares estimator
OpenDSS	Open Distribution System Simulator™
OWA	Optimally weighted average data imputation method
PE	Parameter estimation
PF	Power factor
PMU	Phasor measurement unit
PT	Potential transformer
pu	Per-unit value
PV	Photovoltaic
RMSE	Root mean squared error
SCADA	Supervisory control and data acquisition
SLRPE	Simplified linear regression parameter estimation algorithm
SQL	Structured query language
STLF	Short-term load forecasting
TE	Topology estimation
TSPE	Transmission system parameter estimation
UTC	Universal coordinate time
VAr	Volt-ampere reactive
VSTLF	Very short-term load forecasting
VVO	Volt/VAr optimization
WA	Weighted average
WN	Weeknum defined as a function of WD, HH, and MM

SUMMARY

Efficient control and operation of distributed energy resources (DER) with advanced distribution management system functions such as Volt/VAr optimization (VVO) and conservation voltage reduction (CVR) requires accurate and reliable distribution system modeling, monitoring, and coordination. However, with the large number of parameters and possible load conditions in a distribution system model, there is a high degree of uncertainty with respect to the accuracy and quality of current utility distribution system models. Improving the accuracy and detail of the feeder models is fundamental for advanced distribution management system (ADMS) functions with high penetrations of DERs. Since many DERs are located in the low-voltage secondary circuits, it is becoming important to include the secondary circuits into the distribution models. This is particularly important since the low-voltage secondary circuits have higher per unit impedances, which result in a large share of the feeder per unit voltage drop, as well as some losses.

The feeder models can be enhanced by exploiting novel Big Data from modern distribution system measurement sources such as advanced metering infrastructure (AMI) and photovoltaics (PV) micro-inverters. While the modern distribution measurement sources provide more data, they typically are subject to longer delays and have lower measurement granularity, accuracy, and reliability than transmission system supervisory control and data acquisition (SCADA). Accurate and robust use of the modern distribution system measurements will be a cornerstone of the future advanced distribution management systems.

This dissertation presents accurate, flexible, and computationally efficient parameter and topology estimation methods to calibrate existing utility distribution system

secondary circuit models. Specifically, four algorithms were developed to handle known and unknown secondary circuit topologies with different sets of available measurements. The developed linear regression parameter estimation (LRPE) algorithm is intended for estimating the secondary circuit parameters with fully available AMI data sets when the secondary circuit topology is known. The LRPE algorithm is also extended to handle some meters not reporting voltage measurements. The developed simplified linear regression parameter estimation (SLRPE) algorithm is intended for generating simplified secondary circuit models when only limited measurements are available and the secondary circuit topology is unknown. Finally, the developed linear regression parameter and topology estimation (LRTE) algorithm is intended for estimating the secondary circuit parameters and topology with fully available AMI measurement data sets when the secondary circuit topology is unknown.

The developed parameter and topology estimation methods leverage the Big Data generated by the modern distribution system measurements from AMI and PV micro-inverters. The methods can be run separately for each secondary circuit and can be solved in seconds, despite utilizing thousands of measurement samples to counteract the accuracy and granularity issues related to the modern measurements. The methods are proven to be efficient with the Georgia Tech distribution system with SCADA and AMI measurements, and using large utility feeder models with SCADA and PV micro-inverter measurements. The methods can be utilized for any distribution system model independent of its size or type, and do not require modifying existing utility software.

This dissertation also presents data validation and imputation methods to manage the granularity, accuracy, and reliability issues related to the modern distribution system measurements. The developed data validation methods were effectively used to detect numerous issues in Georgia Tech AMI. Compared to conventional approaches, the developed data imputation method has a superior average accuracy in imputing Georgia

Tech AMI measurements. The method creates a series of imputed samples that have a continuous profile with respect to the adjacent available measurements, which is a highly desirable feature for time-series analyses. The weight parameter of the developed imputation method is trained offline. Using the trained weight parameter, the method is computationally and data efficient and suitable for both offline and online applications.

This dissertation addresses the need for utilities to improve the analytical and operational distribution system modeling accuracy and to manage the Big Data from modern distribution system measurement sources for future advanced distribution management system functions and ubiquitous distributed energy resources. This dissertation also presents several new use cases for the data from AMI and other modern distribution sensors, which encourages further investments into distribution system situational awareness.

CHAPTER 1. INTRODUCTION

1.1 Operational Requirements of Future Smart Distribution Systems

Unprecedented changes are taking place in the electricity distribution systems. Distributed energy resources (DER) such as renewable energy sources, electric vehicles, controllable loads, and electric storage are projected to reach considerable market shares in the future. Driven by falling costs and government incentive programs, solar photovoltaics (PV), currently the most important type of DER at the distribution level, has experienced exponential growth rates, as shown by the International Energy Agency (IEA) statistics in Figure 1. Assuming strong PV growth rates continue, IEA high renewable (hi-Ren) scenario expects PV to make up to 16% of the global electricity supply by 2050 [1].

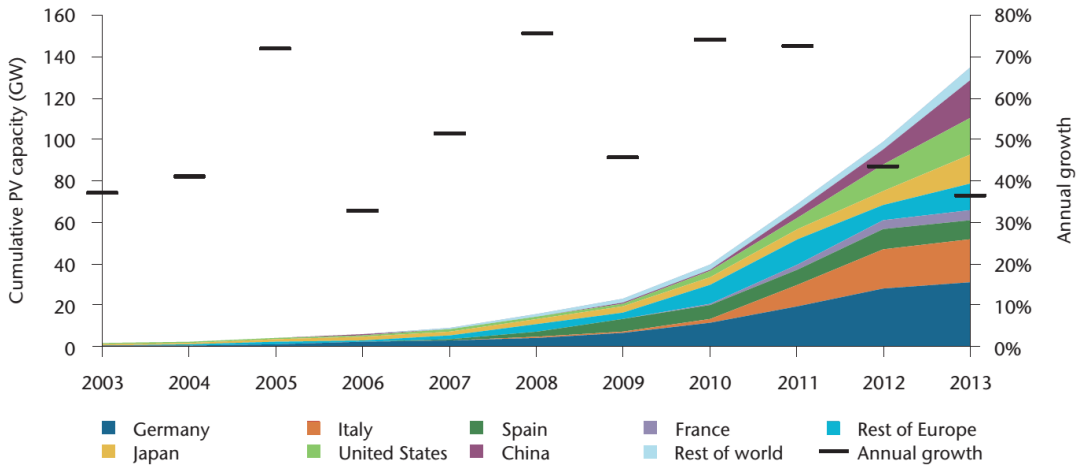

Figure 1. Global cumulative growth of PV capacity [1]

The total U.S. PV capacity has been growing with an increasing rate over the past few years. In 2014 , the U.S. PV capacity grew by 6.2 GW, growth 30% higher than in 2013 and over 12 times higher than five years earlier [2]. With the expected annual growth rate of 6.8% between 2013 and 2040, PV is expected to expand faster than any other source of renewable energy [3, p. 81]. High PV penetration levels are already seen in some geographic areas with high solar resources and supportive politics. In Hawaii, the national

leader of customer PV penetration, there are already circuits where the installed PV capacity exceeds 75% of the daytime peak load and 250% the daytime minimum load, as shown in Figure 2. California leads the nation in both number of PV installations with over 230,000 and with total installed PV capacity at almost 10 GW [4], [5]. An increasing share of PV will be located in distribution systems where it raises concerns of maintaining feeder operating within component loading and voltage standard limits [6]–[10].

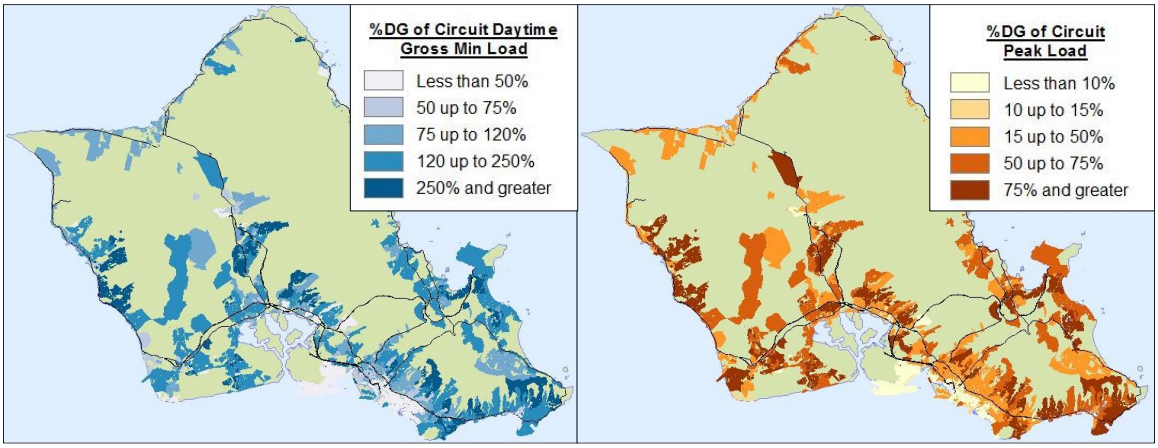

Figure 2. O'ahu island Hawaii PV penetration of circuit daytime peak load (left) and minimum load (right) [11]

Next to renewable energy sources, EVs are projected to have exponential near-term growth rates by multiple entities such as the IEA in Figure 3. The charging of a large number of EVs can introduce significant local load that can result in component overloads and low voltage levels, especially in weak residential distribution systems [12]–[14]. The higher loading can degrade components and may reduce the expected lifetime of distribution equipment, especially the service transformers [14]–[16]. To avoid expensive distribution systems capacity upgrades, controlled or coordinated charging may become necessary [16]–[18].

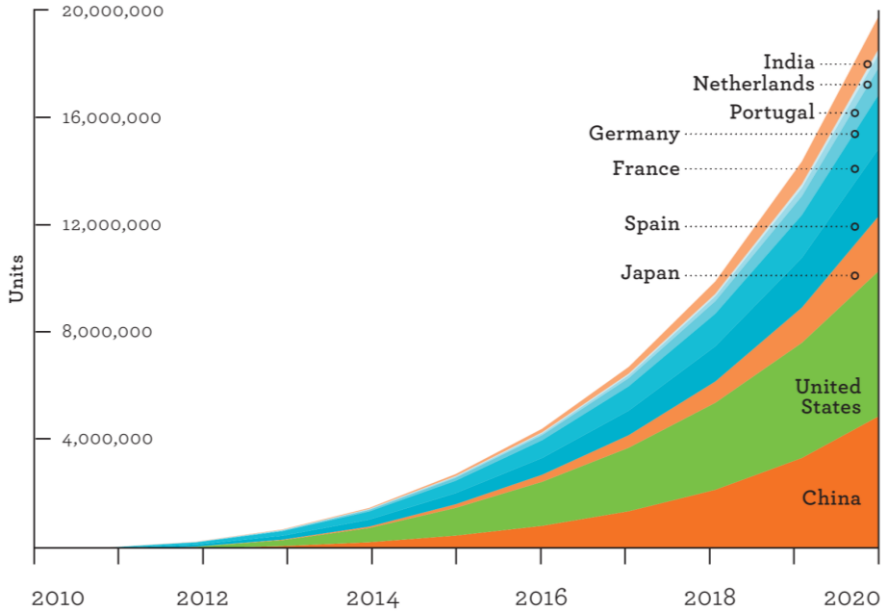

Figure 3. Electric vehicle stock targets before 2020 [19]

In order to maintain economic, high-quality, reliable, and safe distribution system operation under pervasive DERs, faster and more accurate monitoring, coordination and control is imperative [20]–[22]. Much of the distribution system operation is model-based, which means that the models and the input data to those models must be very accurate. Currently, neither is. Instead, the models are often outdated and inaccurate, and measurement data is typically not properly integrated to be fully leveraged. However, emerging data and new sensors have the potential to provide enough information to support the new operational needs [9], [22]–[25].

Simultaneously, the raising customer expectations for service reliability, power quality, and resiliency for natural disasters and other threats are leading to the deployment advanced distribution management systems (ADMS) [22], [26]. The functions of ADMS include fault location, isolation, and service restoration (FLISR); conservation voltage reduction (CVR); peak demand management or demand response (DR); and Volt/VAr optimization (VVO). ADMS functions require more accurate and reliable situational awareness [18], [20], [27]–[29], which is motivating an increase in modern distribution

system measurements such as smart meters and PV micro-inverters [23]. Accurate and robust use of all available measurement information, as well as accurate distribution system models, will be essential for future smart distribution systems with ubiquitous DERs and advanced distribution automation functions [30]–[32].

1.2 Emerging Distribution System Measurements

There is an on-going rapid deployment of modern sensors in distribution systems. Currently, smart meters are the most prevalent type of these modern sensors. There are several smart meter technologies with different capabilities. The combination of electronic meters with two-way communications technology for information passing, monitoring, and control is commonly referred to as advanced metering infrastructure (AMI). Former systems, which have one-way communications for collecting measurement data, are commonly referred to as advanced meter reading systems (AMR) [33]. As illustrated in Figure 4, modern AMI has many functionalities and benefactors compared to the former AMR Plus and AMR.

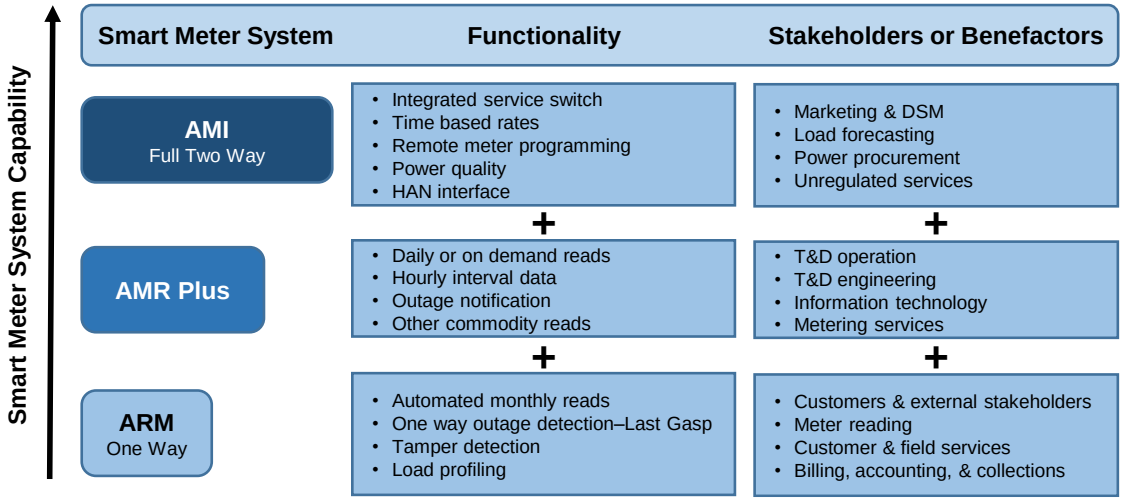

Figure 4. Evolution from AMR to AMI [33, p. 7]

Driven by the additional functionalities and benefactors, the U.S. power industry is clearly shifting to AMI over AMR (Plus) as illustrated in Figure 5. In 2013 for the first time, the number of AMI meters (meters operating two way) exceeded the number AMR meters [34]. At the end of 2013, 35.5% of all U.S. electrical customers had smart meters (AMI or AMR) with the deployment greatly varying by state as illustrated in Figure 6. With 5 (6) states having AMI (AMR) rate over 80% at the end of 2013, the trend is clearly towards a 100% penetration of AMI (two way) meters. For simplicity, the rest of this dissertation refers to all meters that can provide (sub) hourly measurements as smart meters and the corresponding metering systems as AMI.

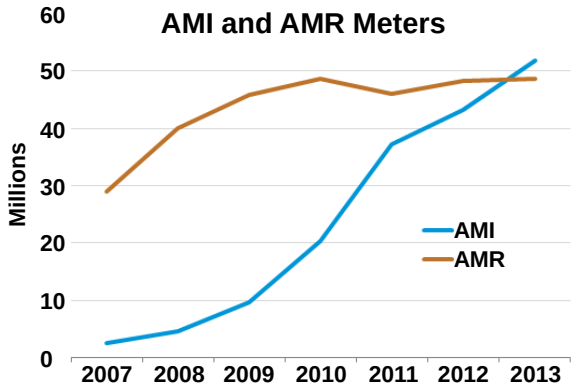

Figure 5. Number of AMR and AMI capable meters in the U.S. [34]

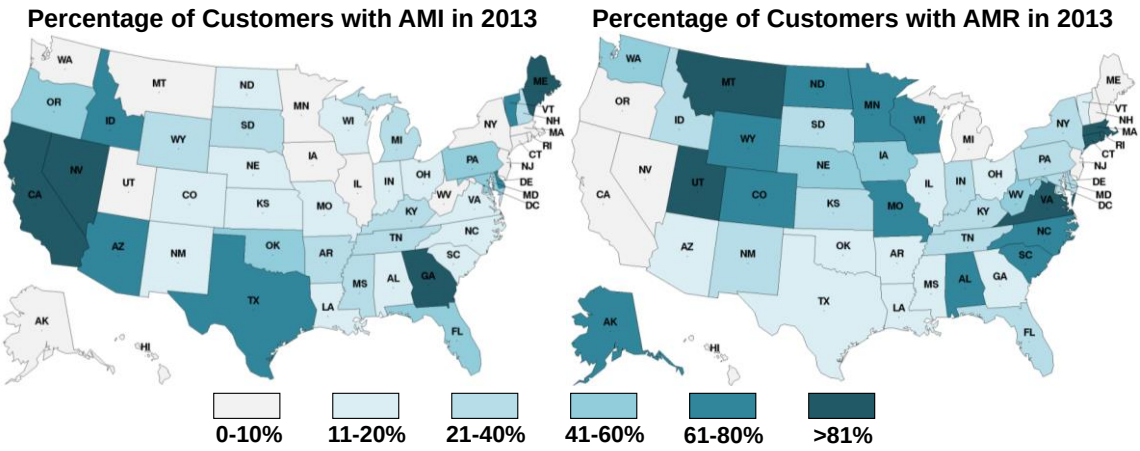

Figure 6. The percentage of AMR and AMI meters in different states in the U.S. [34]

Next to smart meters, many DER sensors also provide valuable measurement data that currently, has not been fully leveraged. The emerging Big Data from smart meters and DER sensors enables a new level of detail for distribution system modeling, analysis, and operation. However, in order to tap the full potential of this Big Data, it must be efficiently and accurately managed.

Smart meters and their instrument transformers must meet national standards such as American National Standards Institute (ANSI) C12.1-2008, C12.10-2011, and C12.20-2010 as well as the IEEE Standard C57.13-2008 with respect to their operation and accuracy [33], [35]–[39]. Therefore, when properly installed and operational, smart meter accuracy tends to be very high – in practice typically within +/- 0.5% or better [33]. The power industry is moving towards even more stringent accuracy standards. ANSI standard C12.20-2010 introduced meter accuracy classes .2 and .5 that correspond to accuracies within +/- 0.2% and +/- 0.5%, respectively [36]. Moreover, a standard for class .1 that corresponds to +/- 0.1% accuracy has been proposed lately [40].

Although smart meters as a group are accurate devices, individual meters can sometimes fail [33]. Individual meters can also be subject to various types of gross errors, many of which are related to meter setup and configuration [41]. Typical errors include improper meter time synchronization, inconsistently set units, and incorrect current transformer (CT) and power transformer (PT) selection, installation, and setup [42].

Compared to supervisory control and data acquisition (SCADA) measurements, modern distribution system measurements typically have lower reliability and granularity as well as longer delays. SCADA is typically very reliable, but a typical well-run large-scale AMI may miss up to 4 percent (1 percent) of hourly power (daily energy) measurement data [41]. Moreover, while SCADA measurements are typically sent to the EMS every 2-10 seconds [29], the most common AMI data granularities are 15-min, 30-min, and 60-min [43], [44]. Smart meter measurements may also be transmitted to the

meter data management system (MDMS) only every few hours or once a day to reduce network traffic [39].

Even with the modern sensors, distribution systems have much lower measurement redundancy than transmission systems. The limited measurement redundancy makes it challenging to detect bad data and to estimate bad/missing measurement samples. Due to the limited measurements, many distribution system sections would be unobservable in state estimation even with the modern sensors [45]. Distribution system state estimator – once deployed – will require accurate and robust use of all the available measurements.

Technically, modern measurement sources could have SCADA-level measurement accuracy, granularity, and reliability, but this would imply unacceptably high costs. Leveraging the Big Data from modern distribution system measurement sources requires robust and efficient ways to detect and manage missing and inaccurate measurements.

1.3 Estimation Methods in Distribution Systems

Much of the distribution system analysis and operation is based on the assumption that the models used to run steady-state simulation are accurate. Circuit models, including the parameter values and topology, may be incorrect as a result of data entry errors, inaccurate equipment data, network changes (e.g. phase balancing), incorrect tap information, etc. [46]. The most commonly encountered errors in the distribution system Geographical Information System (GIS) and power flow models include incorrect component parameters, customers modelled connected to the wrong distribution transformer, and distribution transformers modelled on the wrong phases of feeders [26], [47]. Improving the accuracy of feeder models becomes critical as the numbers of DERs increases because DERs make it more challenging to operate the feeder within ANSI limits. Improved feeder models are also needed for ADMS functions such as Volt/VAr optimization, FLISR, and CVR schemes [26].

It is particularly important to improve the models of the distribution system secondary (low voltage) networks where a large share of the new controllable devices, such as EVs, PV with smart inverters, and demand response, are located [9], [48], [49]. The secondary networks are typically either not modeled at all or modeled with a low level of detail although a significant portion of per-unit voltage drop/raise occurs over the high impedance service transformers and low voltage lines with high losses [9], [48], [49]. With the today's computational power, modeling secondary circuits is not only possible, but it is also necessary to implement useful Smart Grids at the distribution level [49].

The typical approaches to correct circuit errors, such as performing physical inspections or utilizing added measurements, require considerable man hours and additional resources and thus, are not cost-effective [26], [47]. Physical inspections can also be hard to perform in densely populated urban areas with wiring underground and in buildings. There is a growing need for automated procedures to improve the accuracy of distribution system including secondary circuit models with minimal physical inspections.

1.4 Research Objectives

The objective of this dissertation research is to address the need for utilities to improve the analytical and operational distribution system modeling accuracy as well as to manage the Big Data from modern distribution system measurement sources for future advanced distribution management system functions and ubiquitous distributed energy resources.

The research focuses on improving the accuracy and detail of feeder secondary (low-voltage) circuit models. The objective is to develop parameter and topology estimation methods to calibrate existing utility feeder secondary circuit models. The methods must be able to handle known and unknown secondary circuit topologies and different sets of available power and voltage measurements.

The research also focuses on developing data validation and imputation methods to manage the granularity, accuracy, and reliability issues related to the modern distribution system measurements.

The objective is to develop methods that are computationally efficient, easy to implement, and straightforward to integrate with current utility information systems. The developed methods shall also be cost-effective compared to typical approaches through minimizing the need for manual intervention and inspections, and requiring only readily available distribution system information and measurements.

1.5 Dissertation Outline

Chapter 2 presents a literature survey on leveraging modern distribution system measurements for enhanced distribution system situational awareness. First, current industry practices for smart meter data validation are presented. Then, the chapter discusses typical approaches for imputing/estimating bad and missing measurement samples. The chapter also presents the literature on transmission and distribution system parameter and topology estimation methods.

Chapter 3 proposes methods for managing distribution system sensor data. First, the chapter presents the model of Georgia Tech distribution system that includes extensive AMI, making it a perfect case study for smart meter data management and model calibration with parameter and topology estimation. Then, the chapter presents methods for validating smart meter measurements and demonstrates them on the Georgia Tech smart meter measurements. Next, the chapter proposes a novel method for imputing bad and missing smart meter measurements and shows it for the Georgia Tech data. The chapter also presents a method for verifying distribution models with AMI data.

Chapter 4 proposes a linear regression parameter estimation algorithm with fully available AMI data for estimating the parameters of distribution secondary circuits when

the topology is known. The optimal linear regression models for parameter estimation without and with measurement error are presented. The chapter also shows a detailed study on the parameter estimation accuracy with respect to sample size and measurement error level. The parameter estimation results are shown for a three-phase test circuit and for one of the Georgia Tech distribution system feeders.

Chapter 5 extends the secondary circuit parameter estimation scope to cases where the available measurement data is limited. First, the chapter presents a method for handling meters that do not report voltage measurements. Then, the chapter proposes a parameter estimation method to automatically generate secondary circuit models in the case when a utility does not have a dense network of smart meters (or other sensors) in the secondary circuits. The developed results are demonstrated on a test circuit and three large U.S. utility feeder models.

Chapter 6 proposes a joint secondary circuit parameter and topology estimation method for cases when the secondary circuit topologies are not known. First, the chapter demonstrates that exhaustive search of all topologies is impractical. Then, the chapter presents a method for validating an estimated topology in test cases where the true topology is known. Next, the joint topology and parameter estimation method is presented. The method is shown for a test circuit and for the secondary circuits of a Georgia Tech feeder. The chapter concludes by discussing the path towards practical implementation of distribution system topology and parameter estimation.

Chapter 7 presents a summary of the key results and conclusions of the work presented in this dissertation. Chapter 7 also lists the contributions and proposes areas for future work.

CHAPTER 2. LITERATURE SURVEY

This chapter provides a survey of the current industry practices on modern distribution system sensor data validation and imputation. The focus is on smart meter data validation and imputation but many of the discussed principles also apply for the other modern sensors. This chapter also discusses the literature on distribution system parameter and topology estimation.

2.1 Sensor Data Quality and Verification

Incoming measurement data management process is typically divided into two parts: data validation and estimation of missing and inaccurate data [50]. In statistics, data estimation is often referred to as data imputation. The objective of the data validation process is to identify whether the data correctly represents the measured situation. Following the data validation, the data imputation process estimates the identified bad and missing data. Next, literature on data validation and imputation is discussed.

2.1.1 Data Validation

Various reasons can cause inaccurate measurement data. In many cases, measurement data errors can be identified with straightforward checks such as some of the industry best practices listed in Table 1. However, there are other issues, .e.g., incorrect CT ratio that no simple checks can detect. More sophisticated smart meter data validation algorithms, such as [51], are proposed in the literature, but these methods typically either require active user involvement or work only under restrictive load characteristics such as strong daily or long-term load patterns. Unfortunately, load profiles of individual residential customer do not have such characteristics. In the future, data validation or bad

data detection may be incorporated into DSSE allowing the detection to leverage the measurement redundancy of multiple meters. However, most utilities have no DSSE today.

Table 1. Some industry best practices for smart meter interval data validation [41], [50]

Method	Principle
Meter identification	Verify that meter identification is what is expected
Time tolerance	Check meter time tolerance to minimize and correct meter clock drift
Pulse overflow check	Inspect if pulse overflow has occurred at any time interval
Test mode check	Check if meter is left in the test mode
Sum check	Compare the sum of energy consumption over time intervals to the difference in meter energy register values
Spike check	Identify intervals with suspiciously high usage
Reactive channel check	Find intervals with reactive power but without active power consumption
High/low check	Identify suspiciously high or low usage compared to historical values
False zero values check	Verify if zero values are accurate or indicate bad data
Mis-programmed device check	Check that meter parameters are set correctly
Improper meter resets	Verify if any lower register values follow higher ones

2.1.2 Data Imputation

The measurement samples that have failed the data validation process must be imputed. Next, the statistical literature on data imputation is first discussed after which a literature survey on smart meter data imputation is presented.

2.1.2.1 Statistical Perspective on Data Imputation

Missing data handling is a well-established area in statistics. Statistical types of missing data and methods for handling missing data in regression analysis are discussed in [52]–[54]. The conventional methods to handle missing data in regression analysis, such as listwise deletion, pairwise deletion, and dummy-variable adjustment ignore missing data entries completely or consider them by adding dummy variables in regression analysis.

Alternatively, the missing data can be filled in with data imputation [53]. The common data imputation methods are categorized as single imputation, multiple imputation, and maximum likelihood estimation (MLE) methods [52], [53], [55]. Single

imputation methods fill in precisely one value for each missing one whereas multiple imputation methods generate multiple values for each missing entry to better reflect the uncertainty of the missing data. Single imputation methods are the most commonly used approaches to fill in missing values. Single imputation methods, such as replacing the missing values by the mean of existing values or using linear regression to estimate the values, are simple to implement but can lead to biased estimates of certain parameters in statistical modeling such as linear regression. Variance is typically underestimated leading to biased estimates of any parameters that depend on the variance. The standard errors will also be underestimated resulting in inflated test statistics. The larger the fraction of missing data is, the more severe these negative impacts become. [52], [53].

Compared to single imputation methods, multiple imputation and MLE have better statistical properties but have other disadvantages. MLE requires constructing a parametric model for the joint distribution of all the variables with missing data. Constructing such a model can be hard and the results can be sensitive to model choice. Multiple imputation gives different results every time it is run. Both MLE and multiple imputation much more computationally intensive than the single imputation methods.

Imputation of missing air quality data has been analyzed in [55]. This paper compared various single and multiple imputation methods including different interpolation methods, regression-based imputation, multivariate nearest neighbor, self-organizing maps, multi-layer back-propagation nets, and hybrids of these methods. The authors concluded that simple interpolation methods work well for short data gaps but the maximum gap length depends on the characteristics of the imputed variable. The authors also concluded that the imputation performance depends on the amount and characteristics of the missing data. Finally, the authors underlined the importance of using several performance metrics because no single metric can fully capture the imputation performance.

Single imputation and model-based imputation methods are also discussed in [56]. The authors analyze regression imputation but judge it to be computationally too intensive. The authors also discuss MLE and multiple imputation, and conclude that while they are more accurate, they are also much more computationally intensive.

An ensemble algorithm for missing data imputation has been proposed in [57]. The authors propose a bootstrap sampling of prediction interval to estimate the standard error. Unfortunately, while bootstrapping can be very effective in estimating the statistical properties of the predicted values, it is computationally very intensive.

2.1.2.2 Smart Meter Data Imputation

Smart meter data imputation is related to the very mature research field of (electric) load forecasting. Smart meter data imputation is typically applied for relatively short missing data intervals from few hours to a day or two. Thus, smart meter data imputation is mainly related to short-term load forecasting (STLF) and very short-term load forecasting (VSTLF) [58]–[60]. However, STLF and VSTLF research typically focuses on forecasting the total system load [59], [60] that due to aggregation impacts [61], considerably differs from smart meter data imputation that analyzes individual loads or customers.

A power industry best practice to estimate bad/missing smart meter data intervals shorter than two hours is to directly apply linear interpolation to the surrounding data [50]. Bad/missing samples cannot be utilized as end points for the interpolation. Moreover, in some cases, the samples adjacent to the missing samples are not used for interpolation because they may be influenced by the bad/missing samples.

For periods longer than two hours, the typical approach is to construct daily load profiles based on previously validated historical data of “like weekdays” and “like days” [50]. Holidays and other special cases are typically considered separately. Bad/missing

historical samples cannot be utilized for imputation. Moreover, samples for times of power failure should not be used for imputation. Typically, the average daily profile is calculated using three selected days from either the past 90-day period or the previous year. If no data is available during like times, it is customary to use generic customer class load profile that can be scaled to the past month of use. [50].

The historical data imputation approach also has similarities with distribution system load modeling before the advent of smart meters when load measurement data was scarce and statistical models based on load surveys and customer types were used for load modeling [62]. Later, it has been questioned whether the traditional load profiles anymore reflect the electricity consumption patterns [63].

Advanced load profile modeling with smart meter data has been extensively studied in the recent literature. Auto-regressive load estimation model for load monitoring purposes has been shown in [64]. However, the presented model appears to be computationally intensive and does not consider individual customer behavior, e.g., on specific days. Linear Gaussian mixture models and factor analysis was utilized for modeling load profiles in [65]. The authors demonstrated the superiority of their approach over the standard load profiles used in the United Kingdom. In [66], domestic load characteristics such as load peaks and base usage are identified with Bayesian estimation utilizing AMI data with 30-min granularity. Modeling loads as pseudo-measurement through Gaussian Mixture Model is discussed in DSSE [67].

Customer type specific load profiles are computationally much more efficient than imputation and forecasting each customer load separately based on the customer historical smart meter data. However, dividing customers in different segments and utilizing an average load profile for each customer segment clearly ignores any customer specific load behaviors and does not account for spatial load characteristics, such as a load that tends to be higher in certain distribution system area.

An efficient and mathematically rigorous algorithm for power system load data cleaning has been discussed in [68]. The paper however, assumes a linear dependence between individual meter measurements that may be an unrealistic assumption for residential (and other) customer AMI measurements that typically have highly stochastic characteristics. Moreover, the presented approach requires additional communication between smart meters that is typically not available today.

Leveraging smart meter data for individual customer load modeling has also been studied in the literature. In [69], the authors have proposed a time-variant customer specific ZIP-load model and presented an approach for estimating the time-varying coefficients using machine learning methods. Although the presented approach can provide more detail for distribution system analysis, it is not intended for interval data estimation.

Distribution system state estimation with smart meter data has been studied in the recent literature [70]–[72]. Some prior work also exists on the load modeling with AMI data for DSSE such as [73], [74]. Furthermore, short-term load forecasting error has been studied [61]. In the future, data imputation may be integrated into the DSSE. However, since most utilities have no DSSE today, data imputation remains as a separate process.

Overall, the statistical literature on data imputation is rich, but it tends to focus on offline settings where computational requirements are far less important than prediction accuracy. Moreover, the current literature tends to either rely on unrealistic assumptions, requires considerable human input, or is computationally intensive. This dissertation presents a data imputation method with the objective improving the imputation accuracy compared to the current data imputation practices. The presented algorithm is computationally and data efficient making it practical for offline and online distribution system scale applications.

2.2 Distribution System Parameter Estimation

Smart meter and other modern distribution system sensor measurements can be utilized to calibrate the distribution system models with parameter and topology estimation. Next, the literature on distribution system parameter estimation is reviewed. The parameter estimation (PE) problem consists of finding the most likely current system topology and component parameters, each of which are typically known with varying levels of accuracy [46], [75], [76]. Because of the exponential number of possible topologies and parameter value combinations, parameter and topology estimation should not be seen as an alternative for good initial system modeling [76], but rather as a way to calibrate and verify the accuracy of existing utility models. Moreover, it is not advisable to estimate parameters whose initial error is smaller than the average measurement error since for a given measurement redundancy level, the estimated parameter errors are proportional to the average measurement error [75]. In worst case the presence of measurement noise can result in replacing rather accurate original parameter values by less accurate estimated values [75].

Line and transformer parameters can be assumed to be time invariant and estimated off-line, whereas load tap changing transformer tap positions change over time and require online PE [75], [77]. The local measurement redundancy and robustness of offline parameter estimation can be increased by utilizing multiple time steps of measurement data that can be selected free of gross and topological errors [75], [77]. Additionally, offline PE requires no modifications to the existing online algorithms [75].

Traditional transmission system parameter estimation methods, which have been studied at least since 1970s, are typically integrated in the state estimation algorithm and are based either on residual sensitivity analysis or augmented state vectors [46], [77]. In the former type, PE is performed after state estimation by utilizing linear sensitivities between the parameter errors and measurement residuals. In the latter type, the typical state

vector is augmented with additional variables that represent suspicious parameters. The augmented state vector methods apply either normal equations or Kalman filter theory. The augmented state vector methods have surpassed the residual methods, which however, are important for identifying suspicious parameters [75], [77].

Topology errors are often easily identified in state estimation, whereas parameter errors are harder to detect and may go unnoticed for longer periods of time [46], [75], [77]. Topology errors cause several normalized residuals to violate a specified threshold in state estimation algorithm. These residuals correspond to measurements close to the topology error. Similar phenomena is observed with gross measurement errors because of the so-called "smearing effect" that can make it challenging to distinguish between topology errors and gross measurement errors [78]. Since erroneous network parameters have a relatively local impact on the state estimation results, parameter estimation can be performed in a local manner. Accurate measurements typically help in identifying parameter errors [75].

Compared to the well-established transmission system parameter estimation, distribution system parameter estimation (DSPE) is subject to a number of different challenges. Multi-phase asymmetric, mostly radial distribution systems with unbalanced loads, low X/R-ratios, and various connections of transformers and loads make distribution system models complicated [9], [27]. Most utilities do not have existing distribution system state estimators and thus, most conventional transmission system parameter estimation approaches are not directly applicable. Finally, the low number and quality of measurements results in a low measurement redundancy or even the lack of observability in certain circuit sections. For these reasons, DSPE has been studied less than transmission system parameter estimation but is becoming possible to carry out by the advent of AMI and other modern distribution system measurement sources. However, it is imperative to

manage the lower redundancy, reliability, accuracy, and granularity of the modern measurements in parameter estimation.

Using AMI data for transformer load management and service transformer phase error detection has been discussed in [79]. However, the presented method assumes that distribution system state estimator is available. Moreover, the authors do not present methods for calibrating or generating models for the secondary circuits. A linear optimization-based method for topology error detection, parameter estimation, and theft identification has been proposed in [80]. The authors neither estimated the reactances nor leveraged the reactive power measurements. Topology error detection regarding smart meter placement in GIS system is introduced in [47]. In [81], the authors assume a known radial network topology and derive a quadratic equation between the smart meter measurements and upstream node voltage. Then, utilizing this equation, the authors estimated branch parameters using a gradient-based approach with the objective of minimizing the variance of voltage estimates from various smart meters. The approach makes no simplifications to the AC power flow equations but results in an optimization problem with quadratic equality constraints that is computationally much more intensive to solve than linearized approaches. In [82], the author presents practical methods for meter phase identification and meter-to-transformer mapping and secondary modeling by applying a voltage drop equation and linear regression with AMI energy and voltage measurements. This dissertation presents accurate, flexible, and computationally efficient methods for estimating the secondary parameters of existing utility feeder models.

2.3 Distribution System Topology Estimation

In transmission systems, topology estimation (TE) has been studied since 1970s whereas distribution system topology estimation (DSTE) is a rather new research area that is motivated by the increasing penetrations of DER and enabled by modern distribution

system measurement sources. In transmission systems, where the substations topology types are typically known, topology estimation typically focuses on detecting topology errors that can be broadly categorized as errors in the status of switching devices and substation configuration errors [46]. Many of the conventional topology error detection approaches either require a residual vector from an existing state estimator or involve some modifications to the existing state estimator algorithm [46], [77], [83].

Due to the limited deployment of state estimators in distribution systems, these methods are not readily available. Moreover, these approaches are not intended for estimating entire circuit topologies. References [84]–[87] propose topology (switch status) detection algorithms, which do not require existing state estimator, but the methods are not intended for estimating entire circuit topologies. Additionally, the methods proposed in [84], [85], [87] require pervasive micro-PMU measurements, which are currently rare in distribution systems. Micro-PMU products are already available (see e.g. [88]), but due to the high price of these devices, it is unlikely that they will be common anytime soon. Instead, it is likely that distribution system situational awareness will rely on the unsynchronized measurements from AMI and DER sensors.

Distribution system parameter and topology estimation has recently received increasing attention. A linear optimization-based method for topology error detection and parameter estimation is proposed in [80]. In this work, the authors do not estimate the circuit reactances or utilize reactive power measurements. An algorithm for detecting incorrect smart meter placement in a GIS is presented in [47], [89]. The authors utilize historical smart meter data to detect neighboring meters based on their voltage correlations and meter depths in the circuit tree based on the voltage magnitudes. However, instead of estimating network component connectivity (and parameters) of entire circuit models, this approach is mainly intended for detecting errors in existing portion of the utility model. Topology detection algorithm for radial balanced 3-phase feeders is discussed in [90]. The

algorithm, which is based on an approximation of the node voltages, relies on a rather restrictive assumptions that all lines (and transformers) have similar X/R ratios and that all system nodes are monitored. In practice, service drop impedances may vary significantly and typically only the leaf nodes of the radial circuit trees have smart meters and/or DER sensors. Practical methods for meter phase identification, meter-to-transformer mapping, and joint parameter and topology estimation are shown in [82].

Chapter 6 of this dissertation further develops the method shown in [82] by allowing the estimation of any radial circuit topology. The proposed method, which is based on linearized voltage drop approximation and linear regression, is computationally efficient and can easily leverage large measurement samples generating an estimated circuit for a practical-sized secondary circuit. The computational time is within seconds even when thousands of measurement samples are utilized.

CHAPTER 3. DISTRIBUTION SYSTEM SENSOR DATA QUALITY AND MODEL VERIFICATION

This chapter presents methods to manage modern distribution system measurement data quality and reliability issues and to leverage the data for model verification. The first subsection introduces the model of the Georgia Tech distribution system that has extensive AMI making it a perfect case study for data imputation and verification as well as distribution system parameter and topology estimation. The second subsection presents a data validation method for smart meter applications that can be used to identify bad data types in AMI datasets. The third subsection presents a novel method for imputing bad and missing smart meter and other sensor data. The fourth subsection presents a model verification method using AMI data. All the methods are shown for the Georgia Tech smart meter data and distribution system. Most of the work covered in this chapter has been published in [42], [91], [92].

3.1 Georgia Tech Distribution System

3.1.1 Overview of the Distribution System

Georgia Tech owns and maintains its own electricity distribution system consisting of fifteen 19.8 kV distribution feeders, all fed from the same 19.8 kV substation, which serve more than 200 buildings. Figure 7 shows the overall layout of the campus system.

The campus electrical metering infrastructure consists of substation SCADA and building metering. The substation SCADA includes voltage and current measurements at both substation transformers (PT & CT) as well as current measurements through a CT in breakers at each of the 15 feeder heads. The campus buildings have extensive instrumentation, including approximately 400 revenue-grade smart meters, to control and monitor electrical and mechanical signals. Each building has a main meter and typically

several submeters, for billable tenant loads or specific areas of interest such as a chiller or PV system. Every 15 minutes, the smart meter measurements are recorded and aggregated into a database that allows for comprehensive analyses. A communication system and database is also available for setting data loggers for the desired meters over selected time periods at high-granularity down to the time scale of seconds.

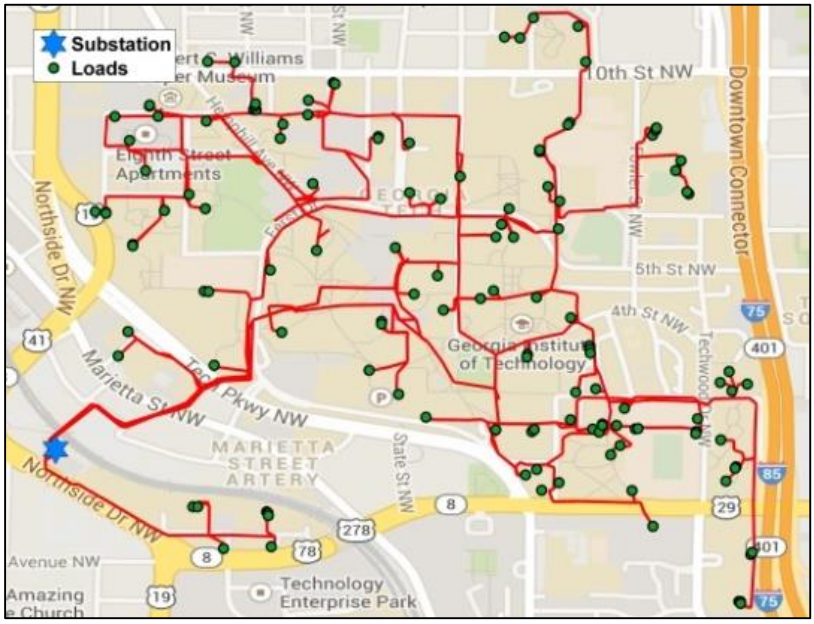

Figure 7. Georgia Tech distribution system

The Georgia Tech system is comprised of a small number of large balanced 3-phase loads. From April 29, 2013 through March 5, 2014, the median and mean of the building average loads (peak loads) were 187 kW (380 kW) and 108 kW (226 kW), respectively. These average and peak load distributions were strongly skewed by six loads at two central campus chiller plants with a total average power consumption of 6.7 MW. The on-campus generation is limited to three rooftop PV units with a total rated capacity of 724 kW_{AC}. Building peak loads over a time period of one year are visualized in Figure 8.

For reliability reasons and to support future campus growth, the system capacity is designed to handle at least twice the peak load, such that a feeder can support maintenance

or outage issues on an adjacent feeder. Thus, in normal conditions, the lines and transformers are not highly loaded, even under peak load. Because of the strong primary system cables, a majority of the voltage drop occurs over the service transformers and secondary circuit lines. The system voltage is controlled solely by the substation on-load tap changers. Most of the system is 3-phase and well balanced. Thus in this system, distribution engineering issues related to single phase laterals and imbalance are not prevalent.

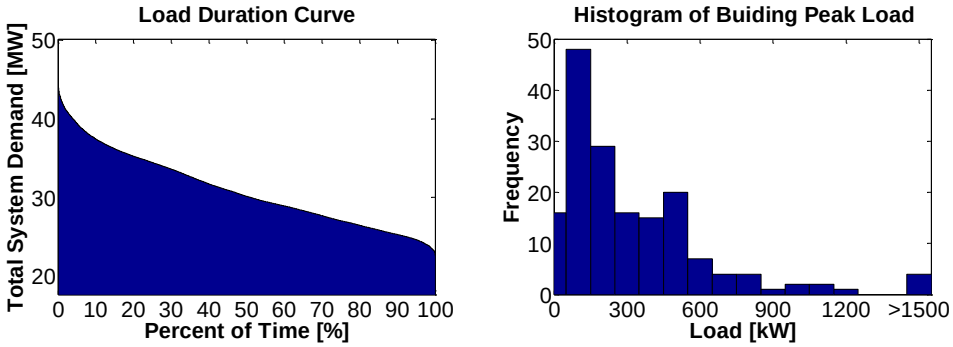

Figure 8. Georgia Tech load duration curve and histogram of building peak load (October 17th, 2012-October 18th, 2013)

3.1.2 Modeling the Distribution System

An accurate OpenDSS model of the Georgia Tech distribution system was implemented with the main purpose of running steady-state 3-phase unbalanced snapshot and time series power flow analyses. The main modeling aspects are briefly discussed next, and the further details can be found in [42], [91].

To simplify the substation model and to allow the two substation buses to be fixed at the measured voltages, the substation (and the upstream transmission system) is simply modeled as two voltage sources behind very small impedances. The voltage sources are set to match the voltage measurements at the buses. The substation model is illustrated in Figure 9.

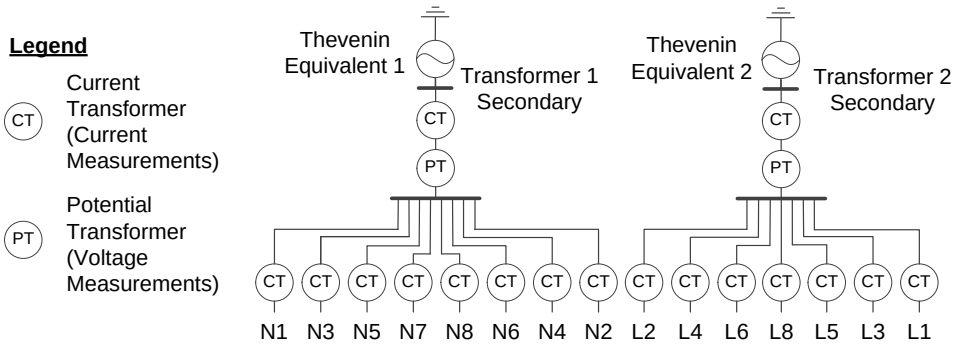

Figure 9. Georgia Tech distribution system substation model

The available Georgia Tech planning documents lists cable insulation classes and materials, conductor type and size, and the grounding sizes. Both overhead lines and underground cables were modeled using the sequence resistances, reactances, and capacitances ($R_1, X_1, C_1, R_0, X_0, C_0$) [93]. By doing this, it was inherently assumed that all lines were symmetric 3-phase lines and un-symmetries of non-transposed lines, unequal phase conductors, non-three-phase lines etc. were ignored [93].

The distribution transformers were modeled with voltage and power ratings, connection (delta-wye), and standard impedance values taken from manufacturers' data. Limited information is readily accessible for the Georgia Tech secondary circuits, i.e., the circuits between the distribution transformers and the smart meters in the buildings. At the same time, an accurate model of the secondary circuits was necessary to verify the accuracy of the simulated voltages to the smart meter data. As a first approximation, each meter was assumed to be connected to its distribution transformer by 30 meters of 500 kcmil copper ethylene propylene rubber (EPR) cable. This principle is illustrated in Figure 10. The secondary system topology and the parameters are refined with parameter estimation discussed in Chapters 4-6.

All loads were modeled as fixed PQ loads since each load represented a smart meter. This allows historical analyses to be simulated with load values fixed to

measurement values, which was a critical part to perform parameter estimation. The distribution transformer, secondary system and load models are illustrated in Figure 10.

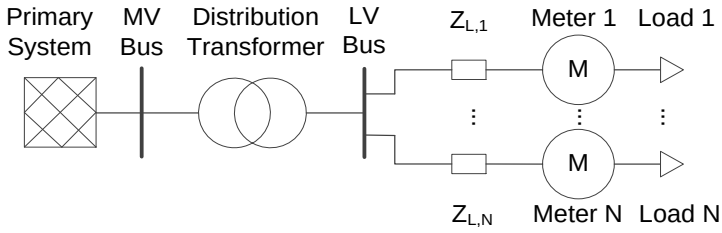

Figure 10. Georgia Tech distribution transformer and secondary circuit model

3.2 Data Validation Methods for Smart Meter Applications

Smart meters are required to meet accuracy and operation standards such as ANSI C12. Their accuracy tends to be high – typically ranging from 0.2 % or 0.5% (class 0.2 or 0.5 meters), to +/- 2 % [33], [37], [39]. However, there are multiple factors that can affect the data accuracy and performance delivered by the Advanced Metering Infrastructure (AMI) and Meter Data Management System (MDMS). Therefore, it becomes imperative to discern if the measured data correctly reflects the events that actually occurred during a time interval of interest.

Large amounts of smart meter data brings along typical data-related issues such as missing and inconsistent values, values stored in the wrong places at wrong times, etc. The data has to be "cleaned" or preprocessed from these inconsistencies before it can be used. Some of the smart meter data issues can be cleaned by inspecting the data of each individual meter whereas other issues can only be detected by comparing the data from one set of sensors with the data from another set of sensors.

3.2.1 Individual Meter Data Validation

The first round of data validation involves identifying problems that can be detected independent of other measurements. Table 2 lists some of the industry best practices for individual smart meter data validation and how many meters were influenced in the

Georgia Tech AMI. Georgia Tech AMI data was validated following these and other steps. Figure 11 illustrates the different issues encountered in the Georgia Tech AMI. It should be noted that some of the issues, e.g., no or zero values, may very well be normal operation of the specific load. However, the issues would raise a flag in meter validation that must be manually inspected. The identified data issues, many of which result from errors in setting up the meters or the data storage, are discussed next. Similar data issues are also expected in other AMI.

Table 2. Identified smart meter data issues in Georgia Tech AMI

Issue	Number of meters influenced in Georgia Tech AMI
Configuration / setup	A meter with wrong CT setup Submeter reported as main meter
Improper time stamp	34 meters with values stored into MDMS a few millisecond off the time stamp 38 meters at wrong time zone
Measurement granularity	10 meters reporting instantaneous values only
Consistency of units	8 meters reporting MW and Mvar instead of kW and kvar
False values	3 meters reporting unrealistic values

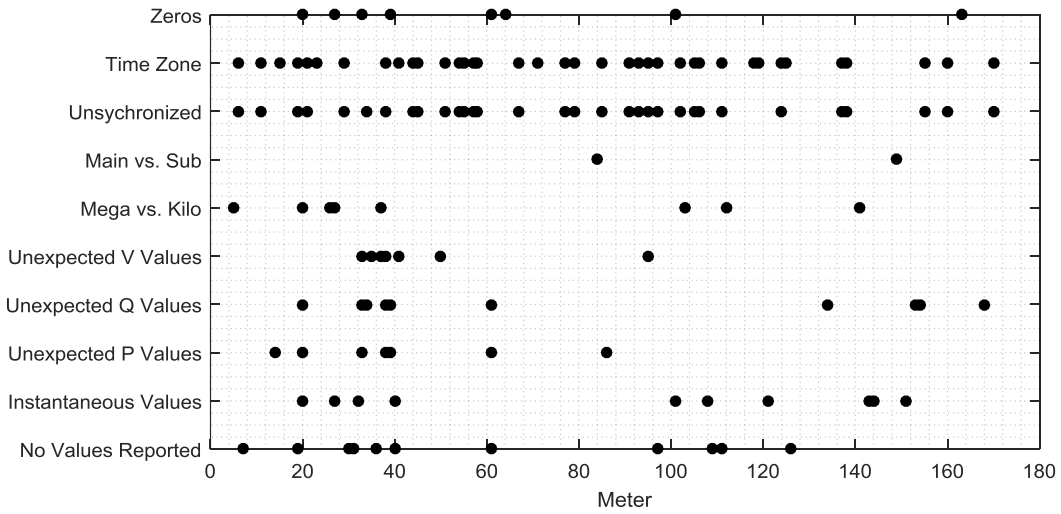

Figure 11. Issues encountered in the Georgia Tech AMI

Configuration: This common source of accuracy error stems from an improper configuration of smart meters and devices. For instance, PTs may be connected line-to-line but programmed as a line-to-neutral reading or vice versa. It may also happen that a wrong

CT ratio has been set for the meter resulting in wrong current measurements. Finally, the single-phase measurements may also be recorded as total 3-phase measurements. These issues may be hard to detect without detailed information of the measured load. In the Georgia Tech system, a thorough inspection of building load characteristics and the time series feeder data validation approach shown in section 3.2.2 revealed a meter with wrong CT setup and a building sub meter identified as one of the main meters.

Improper Timestamp: Georgia Tech smart meters are loosely time synchronized, meaning that they report their measurements at more or less the same time. Prior to storing the data into the database, meters must therefore be synchronized to the Universal Coordinate Time (UTC), and problems may arise when a smart meter is not set in the correct time zone. Data validation revealed a number of meters that were setup in a wrong time zone.

Measurement Granularity: Smart meter data is typically stored with a relatively low time granularity of, e.g., 15, 30, or 60 minutes. Typical ways to represent the interval measurements are either instantaneous values or time-averaged values. When measurement data is used across a set of sensors, the implications of potentially varying time interval representations should be considered. For example, time-averaged values tend to have smoother profiles than instantaneous values. Several Georgia Tech smart meters report instantaneous values influencing, e.g., the time series feeder data validation discussed in the next subsection.

Consistency of Units: In order to draw meaningful conclusions from the data, it is necessary to report the same unit for a particular measurement across all the smart meters. For instance, some load measurements were found to store the values in MW and Mvar even though kW and kvar were expected.

False Values: Values that remain zero or constant for longer periods of time can be an indication of a potential source of inaccurate data. In the Georgia Tech system, some

research and other facilities had very abnormal load behavior making it very challenging to identify when this was a consequence of meter/communication issue and when it was the true load behavior. For constant values, it is important to distinguish between the accuracy of a meter and the numerical accuracy that is used to store those values into the database. For example, a protection CT may have a relatively low accuracy but the values may be stored to SCADA with several digits. As a result, it may seem that the values remain constant over a time period even though the values remained constant only to the accuracy of the protection CT.

Missing Measurement Intervals: It is unlikely that a distribution system will always have all the meters fully operational due to several reasons. For example, it is not unusual to have ongoing renovation work for long periods of time. At a specific time, there might also be meters that are recording and internally storing measurements, but not transmitting them to the database due to a breakdown in the communication. Once communication is restored, the values may or may not show up in the database. A method for imputing the missing measurements is presented in section 3.3.

3.2.2 Time Series Feeder Data Validation

Some measurement data problems may not be immediately obvious by inspecting the data of each meter separately. A powerful method for finding measurement data issues is to compare the sum of the active and reactive powers reported by all the meters on a feeder for a specific time period to the substation SCADA feeder measurements. To do this, the active and reactive power measurements should be available both at the feeder head and at all (or most of) the loads and distributed energy resources along the feeder. This approach is similar to the widely used method of comparing customer energy measurements to the substation measurements but offers far greater detail and potentially reveals more information, such as the profile of unmeasured loads.

Sometimes instead of active and reactive power, only the current magnitude measurements are available from the feeder head breaker CTs. In this case, the comparison above can be done with apparent powers. The measured feeder apparent power at a measurement time instance t is calculated with

$$S_{SCADA,t} = \sqrt{3} V_{LL,t} I_t, \tag{1}$$

where $V_{LL,t}$ and I_t are the substation (feeder) PT line-to-line voltage and feeder CT/breaker current measurements at time instance t, respectively. These values are compared to the aggregated feeder apparent powers $S_{AG,t}$ that are calculated from the aggregated active and reactive power measurements of the meters on the feeder

$$S_{AG,t} = \sqrt{\left(\sum_{i=1}^{N_M} P_{i,t}\right)^2 + \left(\sum_{i=1}^{N_M} Q_{i,t}\right)^2}, \tag{2}$$

where $P_{i,t}$ and $Q_{i,t}$ are meter i active and reactive power measurements at the measurement instance t, respectively and N_M is the number of meters. If the feeder has $N_C \neq 0$ capacitors and/or $N_{DER} \neq 0$ distributed energy resources such as PV, the aggregated apparent power at time instance t is calculated with

$$S_{AG,t} = \sqrt{\left(\sum_{i=1}^{N_M} P_{i,t} + \sum_{j=1}^{N_{DER}} P_{j,t}\right)^2 + \left(\sum_{i=1}^{N_M} Q_{i,t} + \sum_{j=1}^{N_{DER}} Q_{j,t} + \sum_{k=1}^{N_C} Q_{k,t}\right)^2}, \tag{3}$$

where $\sum_{j=1}^{N_{DER}} P_{j,t}$, $\sum_{j=1}^{N_{DER}} Q_{j,t}$, and $\sum_{k=1}^{N_C} Q_{k,t}$ are the aggregated feeder measured/estimated DER active power, DER reactive power, and capacitor reactive power, respectively. This analysis includes a baseline error that reflects unmetered loads/generators, distribution losses, and data inaccuracies but, provided that the fraction of unmetered load/generation is small, average values greater than 5% are usually indicative of the load not being metered or something fundamentally being incorrect with the feeder smart meter data.

In the Georgia Tech system, the aggregated feeder power analysis was done with apparent powers since the feeder heads at the substation have only breaker CTs that do not provide power factor measurements. The voltages are measured only at the LTC secondary

buses. Since the Georgia Tech system had no capacitors and the PVs are behind the smart meters, the aggregated apparent powers were calculated with (2). Figure 12 A) and B) show a distribution feeder whose aggregated load almost perfectly matches the substation measurement, whereas Figure 12 C) and D) present a considerable mismatch. The accuracy was characterized with mean absolute percentage error (MAPE) calculated over the N measurement samples with

$$\text{MAPE} = \frac{1}{N}\sum_{t=1}^{N}\left(\frac{\left|S_{SCADA,t} - S_{AG,t}\right|}{S_{AG,t}}\right). \quad (4)$$

MAPEs are 0.44 % and 5.54 % for the two sets of graphs in Figure 12, respectively. A MAPE value of 5.54 % indicated that a portion of the load on that feeder was not being measured or that multiple meters were not operating properly. A thorough audit of the metering devices and ancillary infrastructure revealed further measurement inconsistencies discussed in section 3.2.1.

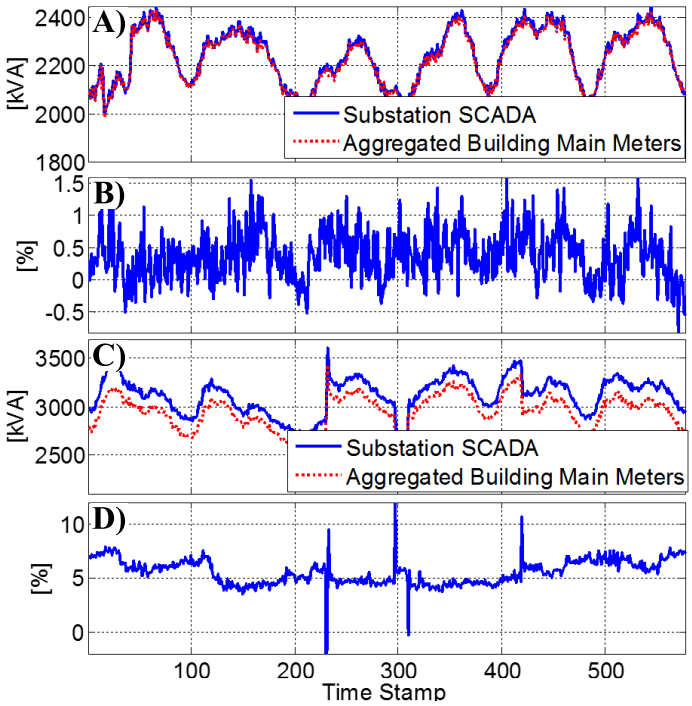

Figure 12. Example of aggregated load comparison (April 27th, 2014 - May 3rd, 2014): feeder without bad data A) kVA difference, B) percentage difference (MAPE=0.44 %), feeder with bad data C) kVA difference, D) percentage difference (MAPE=5.54 %)

Figure 13 shows the percentage errors of all the feeders after detected meter data inconsistencies were fixed. The average and the highest feeder MAPE for the time period from May 28, 2014 through September 17, 2014 were 1.50% and 3.50%, respectively. The error variability in Figure 13 is mainly explained by the two following aspects. First, Georgia Tech campus has several large rapidly changing loads, whose variability is not sufficiently captured by the loosely time-synchronized 15-min average AMI and SCADA data. Second, some of the campus meters report only instantaneous measurement values that tend to be much more variable the 15-min average values thus, creating high spikes seen in Figure 13. Despite the error variability, all the MAPE values are in expected range that can be explained by unmetered campus loads (parking places, street lighting, etc.), network losses, and measurement inaccuracies. Figure 13 underlines the importance of performing the feeder power validation over a measurement time period. Due to the error variability, the information that is received from feeder power validation over a single measurement sample can be limited.

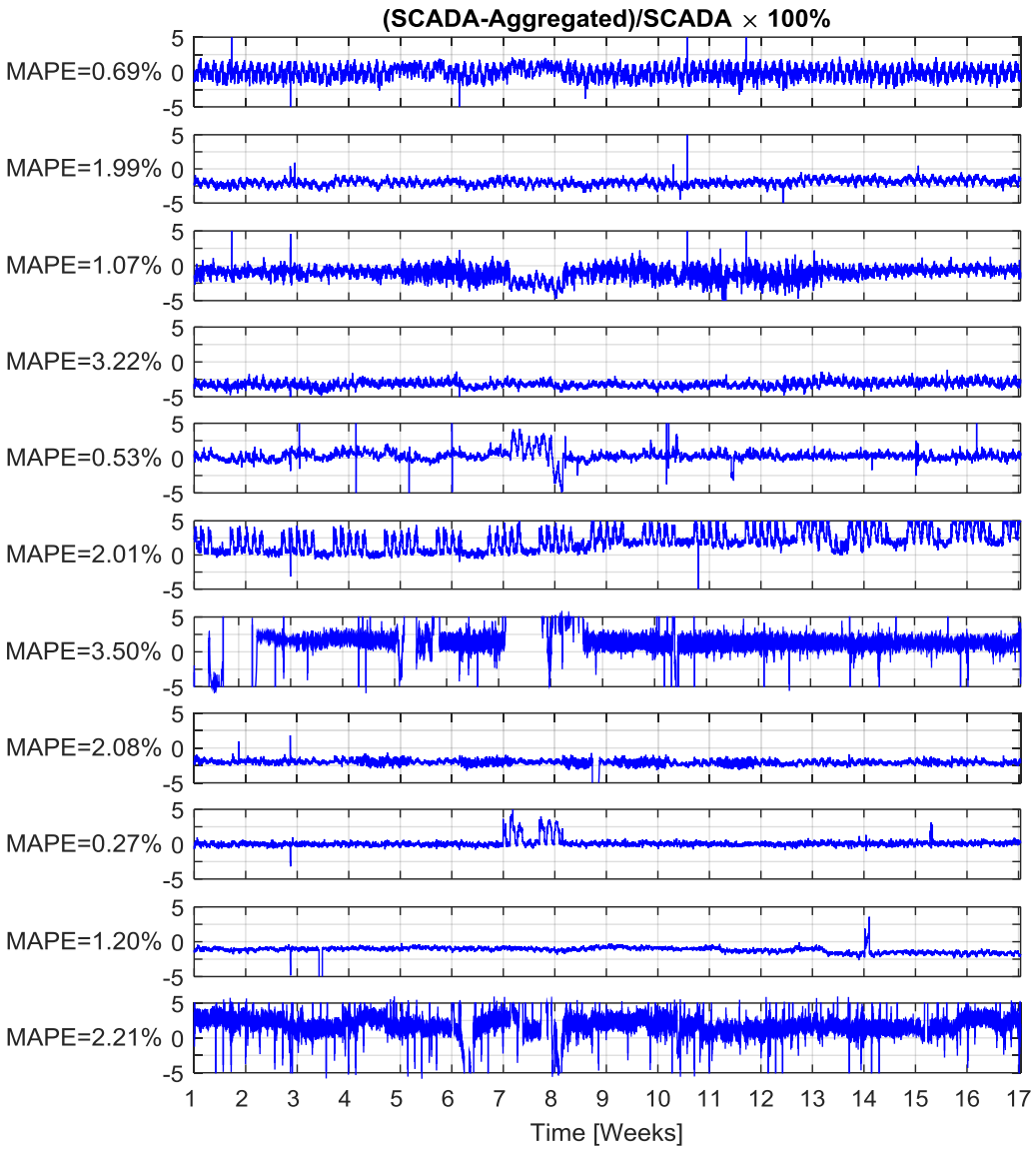

Figure 13. Georgia Tech feeder percentage difference between SCADA measured feeder kVA and aggregated feeder metered kVA from May 28, 2014 through Sep 17, 2014. Plots 1-7 from the top represent a loop feeder and plots 8-11 from the top represent two networked feeders each

3.3 Novel Method for Data Imputation

Once the measurement data has been validated, the identified bad and missing measurement samples need to be estimated or imputed for applications that require full data sets such as the time series power flow analysis. This section presents a novel data imputation method that balances between accuracy and computational and data

requirements. The proposed method is aimed for estimating missing and bad samples in historical smart meter data, or in other situations when only limited computational and data resources can be spared for each missing or bad data interval.

3.3.1 Smart Meter Data Imputation Method

This section presents a computationally and data efficient optimally weighted average (OWA) smart meter data imputation method that is practical for offline and online applications. The method only requires the historical load power measurements from the smart meter (or other sensor). In particular, the method does not require measurement (e.g. customer) specific information or other explanatory variables such as weather.

The proposed smart meter data imputation scheme leverages two typical smart meter data characteristics. First, the data tends to be rather continuous over a short time interval, meaning that short time intervals of missing/bad measurement samples have likely similar characteristics as the adjacent available data. Second, since the data is strongly driven by human consumption patterns, the data tends to have similar characteristics over time periods with similar human activity. For example, the data characteristics of weekdays tend to be different to weekend days, mornings different to evenings, etc.

3.3.1.1 Linear Interpolation Imputation

There are several ways to estimate short intervals of missing samples from the adjacent available samples. Nearest-neighbor and interpolation are particularly commonly used approaches. In the nearest-neighbor approach, the missing samples are simply set equal to the closest available sample or an average of them. Nearest neighbor may work well for a few missing samples but for slightly longer missing data periods, interpolation is preferred since it results in estimates that are continuous with the adjacent available measurements. There are numerous alternative interpolation methods including linear,

cubic, and spline interpolations. The proposed OWA data imputation method utilizes linear interpolation since it tends to have more consistent behavior for missing data with different characteristics compared to cubic, spline, or other more complicated interpolation methods [55].

Linear interpolation (LI) imputation estimates a missing value y_i from the closest preceding and succeeding available values y_h and y_j with

$$\hat{y}_i^{LI} = y_h + \frac{y_j - y_h}{x_j - x_h}(x - x_h), x_h < x_i < x_j. \tag{5}$$

LI imputation is simple to implement, computationally very fast, and only requires two available samples for each missing data period. On the other hand, the accuracy of LI imputation typically decreases as the length of the missing data period increases.

3.3.1.2 Historical Average Imputation

LI imputation tends to perform poorly on long periods of missing data, and better estimates can be derived from representative periods of historical data. The simplest approach to impute missing values with historical data is to use the sample from the previous hour, day, or month. Using a single sample however, can result in highly variable estimates whose accuracy may strongly depend on the missing sample times. The proposed OWA data imputation method utilizes historical average (HA) imputation method that estimates each missing sample y_i as an average of N_H representative historical samples $y_j, j \in \mathcal{H}, |\mathcal{H}| = N_H$

$$\hat{y}_i^{HA} = \frac{1}{N_H}\sum_{j \in \mathcal{H}} y_j. \tag{6}$$

To characterize the set $\mathcal{H}$, we define "weeknum" (WN)

$$WN = WD + \frac{HH}{24} + \frac{MM}{24 \times 60}, \tag{7}$$

as a function of the weekday $WD \in \{1,\ldots,7\}$ (1=Monday, …, 7=Sunday), hour of the day $HH \in \{1, \ldots, 24\}$, and minute of the hour $MM \in \{1, \ldots, 60\}$. Now, the set $\mathcal{H}$ is defined to consist of historical samples whose day of the year (DOY) and WN are within selected spans of the missing sample. In this dissertation, the DOY span of ± 8 days and the WN span of $\pm(\frac{1}{24} + \frac{1}{24\times 60})$ (1 hour and 1 minute) were used. The DOY assures that the historical mean is calculated over samples with similar seasonal characteristics. The WN guarantees that the historical mean is calculated over samples with similar days of the week and times of the day. Holidays and other special days are handled separately or if sufficient data is not available for them, they are categorized as Sundays. This definition of $\mathcal{H}$ results in smooth historical average profiles for sequential missing samples. If "hard" time selection criteria, such as equal season, equal WD, and equal HH was used, the sequential imputed samples would have jumps when the season, weekday, hour, etc. change.

The accuracy of the HA imputation depends on the characteristics of the data and requires clear historically repeating patterns. With these assumptions, on long missing data periods, HA imputation is expected to have a better average performance compared to LI imputation.

3.3.1.3 Optimally Weighted Average Imputation

Next, an optimally weighted average (OWA) imputation method is presented with the objective of leveraging the LI imputation accuracy for short missing data periods and the HA imputation accuracy for longer missing data periods. The OWA imputation estimates a missing data sample y_i as the weighted average of the LI imputed values $\hat{y}_i^{LI}$ and the HA imputed values $\hat{y}_i^{HA}$

$$\hat{y}_i^{OWA} = w_i \hat{y}_i^{LI} + (1 - w_i)\hat{y}_i^{HA}. \tag{8}$$

The (positive) weight parameter w_i is set to exponentially decay with respect to $d_i > 0$, the (positive) distance (in samples) to the closest (preceding or succeeding) available sample

$$w_i = e^{-\alpha d_i}. \tag{9}$$

where α is a (positive) weight parameter. For small d_i (i.e. $w_i \approx 1$), the OWA imputed value $\hat{y}_i^{OWA}$ mainly depends on the LI imputed value $\hat{y}_i^{LI}$. For large d_i (i.e. $w_i \approx 0$), the OWA imputed value $\hat{y}_i^{OWA}$ depends mainly on the HA imputed value $\hat{y}_i^{HA}$. Figure 14 illustrates the weight function w_i dependence on α and d_i. For $\alpha > 2$, the HA imputed values are almost exclusively used for all but the first missing sample. Thus, it is reasonable to restrict $\alpha \in [0,2]$. The optimal value of α depends on the measurement data characteristics including the variability and the historical patterns of the data. The question remains about what value of α to select, so next, a method to optimize α is presented.

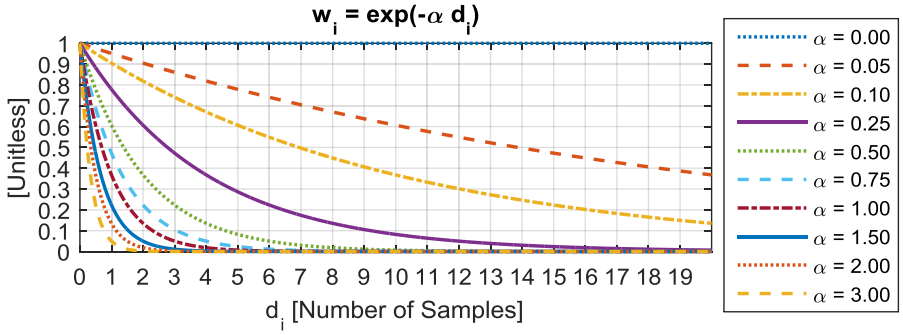

Figure 14. Optimally weighted average imputation weight function dependence on the weight parameter and the distance to the closest available sample

3.3.1.4 Optimal Weight Parameter Selection for a Training Data Period

The optimal weight parameter (for a training data period) α_{opt} minimizes the error $F(\alpha)$ between the imputed samples and the training data samples

$$\alpha_{opt} = \underset{\alpha}{\operatorname{argmin}}\, F(\alpha) = \underset{\alpha}{\operatorname{argmin}} \sum_{i=1}^{N} F_i(\alpha) \tag{10}$$

With squared error, $F_i(\alpha)$ is given by

$$F_i(\alpha) = \left(\hat{y}_i^{OWA} - y_i^{true}\right)^2 = \left(e^{-\alpha d_i}\delta_i^{LH} + \delta_i^{HA}\right)^2 \quad (11)$$

where $\delta_i^{LH} = \hat{y}_i^{LI} - \hat{y}_i^{HA}$ and $\delta_i^{HA} = \hat{y}_i^{HA} - y_i^{true}$. A necessary condition for an optimal solution α_{opt} is that the derivative vanishes $F'(\alpha) = 0$. Such so-called critical points can be found, e.g., with the Newton's method starting at initial value $\alpha = \alpha_0$ and iterating with

$$\alpha_{k+1} = \alpha_k - \frac{F'(\alpha_k)}{F''(\alpha_k)} \quad (12)$$

until a selected convergence criteria is satisfied. The first and the second derivatives of $F(\alpha)$ are given by

$$F'(\alpha) = \sum_{i=1}^{N} -2d_i\left(e^{-2\alpha d_i}(\delta_i^{LH})^2 + e^{-\alpha d_i}\delta_i^{LH}\delta_i^{HA}\right) \quad (13)$$

and

$$F''(\alpha) = \sum_{i=1}^{N} 2d_i^2\left(2e^{-2\alpha d_i}(\delta_i^{LH})^2 + e^{-\alpha d_i}\delta_i^{LH}\delta_i^{HA}\right). \quad (14)$$

If $F(\alpha)$ was a convex function, algorithm (12)-(14) converges to the globally optimal solution of (10) independent of the chosen initial value α_0. However, if $F(\alpha)$ is non-convex, algorithm (12)-(14) may diverge for a poorly chosen initial solution. Unfortunately, the convexity of $F(\alpha)$ may depend on the test samples and imputed samples through δ_i^{LH} and δ_i^{HA} as shown next. For the error function $F(\alpha)$ to be convex, its second derivative must be non-negative $F''(\alpha) \geq 0$. A sufficient condition for this is each of the summands to be non-negative

$$2d_i^2\left(2e^{-2\alpha d_i}(\delta_i^{LH})^2 + e^{-\alpha d_i}\delta_i^{LH}\delta_i^{HA}\right) \geq 0, \forall i. \quad (15)$$

Since, $\alpha \geq 0$ and $d_i \geq 0$, the condition (15) is equal to

$$2e^{-\alpha d_i}(\delta_i^{LH})^2 + \delta_i^{LH}\delta_i^{HA} \geq 0, \quad (16)$$

which indicates that the convexity of $F(\alpha)$ may depend on the test data and the imputed data δ_i^{LH} and δ_i^{HA}. The error function $F(\alpha)$ is nonconvex for any set of training samples

y_i^{true} and imputed samples $\hat{y}_i^{LI}$ and $\hat{y}_i^{HA}$ that result in $F''(\alpha) > 0$. As a result, Newton's method may diverge for a poorly chosen α_0. In practice, good convergence is obtained by selecting a small (but positive) α_0 (e.g., $\alpha_0 = 0.001$). Examples of the shape of $F(\alpha)$ are shown in Figure 15 for different lengths and locations of datasets δ_i^{LH} and δ_i^{HA}. In some cases, $F(\alpha)$ has a clear minimum whereas in others, the function is very flat or even almost constant.

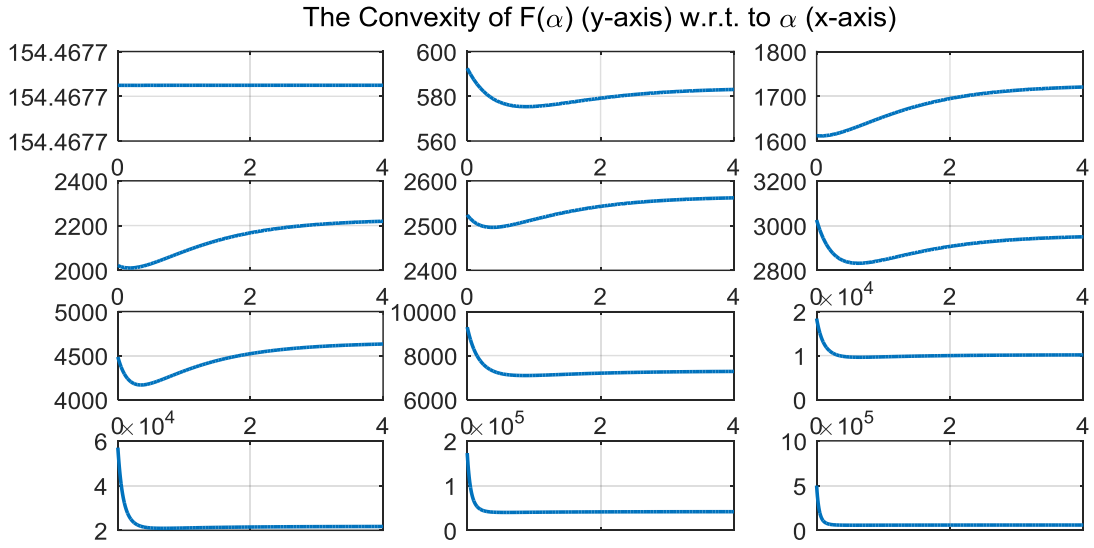

Figure 15. Examples of the error function shape with respect to the weight function parameter for different training data and respective imputed data sets

Figure 16 shows an example of a training data period of 50 samples with the true known values and the values estimated with HA, LI, WA ($\alpha = 0.10$), and OWA imputation. Clearly, for such a long time period, the LI imputation accuracy suffers. Better imputation accuracy is achieved with a linear combination of LI and HA imputation (WA) and best accuracy is obtained with the optimal weight parameter (OWA)

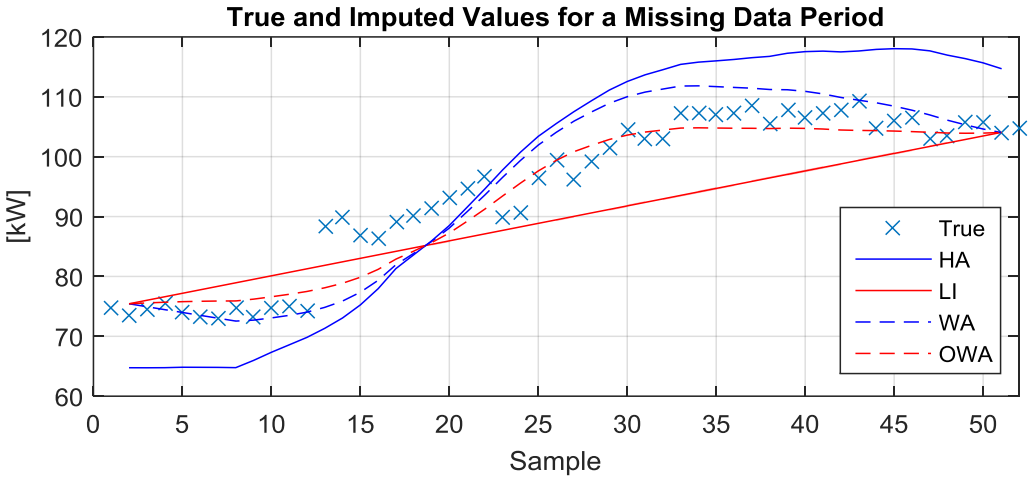

Figure 16. An example of a training data period with the true known values and the imputed values

3.3.1.5 Globally Optimal Weight Parameter

The optimal weight parameter α_{opt} depends on the characteristics and the length of the missing period. Thus, different α_{opt} values are obtained using different training data period characteristics and lengths. The distribution of α_{opt} can be estimated by optimizing α over a set of training data periods with randomly selected lengths and locations. The missing data period lengths can be sampled from known distribution of missing data period lengths (if available). The globally optimal α can be estimated from the mean (or median) of the obtained α_{opt} sample distribution.

Algorithm 1 lists the process of estimating the weight parameter α_{opt} for a meter. The optimal weight parameter α_{opt} of a meter is optimized only once and is stored in the MDMS. Afterwards, missing data is estimated with (8) using the optimized α_{opt}. The results shown in in section 3.3.2 indicate that good estimate of α_{opt} distribution can be obtained with $N_{period} = 100$ for typical missing data period lengths. If dealing with a large number of meters, Algorithm 1 can be executed for a subset of the meters and the mean (or median) of the resulting α_{opt} distribution can be utilized for all meters.

Algorithm 1. Meter weight parameter optimization algorithm

1. Randomly choose the first samples of the training data periods for N_{period} training data periods and $N_{lengths}$ training data period lengths.
2. Construct an array of timestamps of all the samples needed for imputing the training data samples with HA and LI imputation.
3. Fetch the necessary data from the MDMS.
4. Repeat **1.** – **3.** For periods with (true) bad/missing samples.
FOR $N_{lengths}$ training data period lengths
5. For each sample of each training data period, impute the values $\hat{y}_i^{HA}$ and $\hat{y}_i^{LI}$ and calculate δ_i^{LH}, δ_i^{HA}, and d_i.
6. Use (12)-(14) to find α_{opt} that minimizes $F(\alpha)$ over all missing data periods. Store the optimal α_{opt}.
ENDFOR
6. Choose the globally optimal α, e.g., as the mean (or median) of the distribution of α_{opt} values.

3.3.2 Imputation on Georgia Tech AMI Data

Next, the OWA data imputation method is shown for smart meter measurements from the Georgia Tech distribution system. In the Georgia Tech network, the major cause of missing data was ongoing renovation and maintenance work that commonly takes extended periods of time. The data reliability from May 2012 through May 2013 was approximately 97.28%, meaning that 2.72% of the data points were missing. This was largely because of planned outages (maintenance, retired meters, etc.), with only 0.08% of the measurements unexpectedly missing. Similar performance can be expected in other systems of the same type and scale. As seen from these numbers, the main data reliability challenge in such systems is to deal with data that is missing during planned outages that can last for weeks or months. The longer a measurement is missing, the more challenging it can be to accurately estimate its value because of changes in load characteristics, especially for construction or renovation. Alternatively, the possibility of complete disconnection of load also increases.

3.3.2.1 Detailed Analysis for a Georgia Tech Smart Meter

The OWA data imputation was first analyzed with the active power measurements of one of the Georgia Tech smart meters. The analyzed smart meter is located in a building that is mainly dedicated for classroom and office purposes. As a result, the building energy consumption has a clear historical pattern driven by the classroom and office activity as illustrated for a two-week period in 2013 in Figure 17.

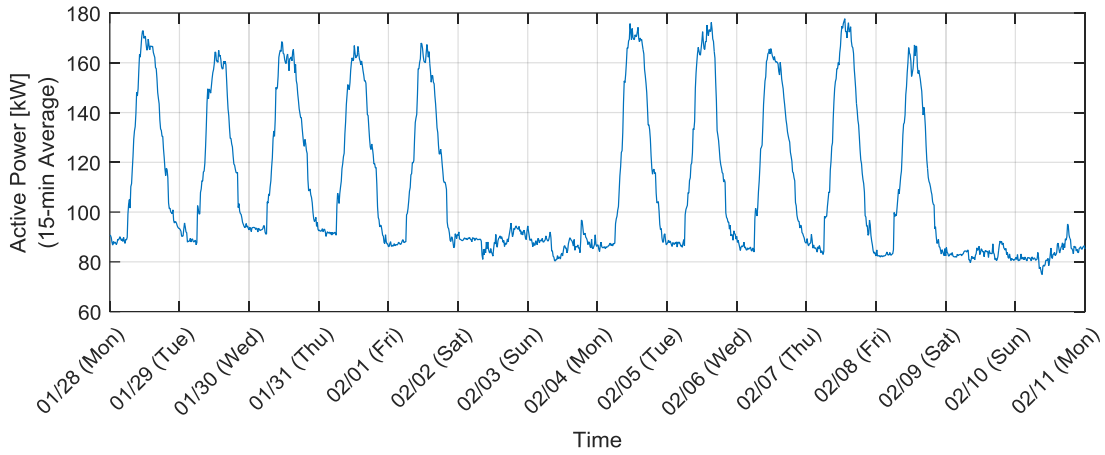

Figure 17. The 15-min average active power measurements for the analyzed Georgia Tech building from January 28, 2013 through February 11, 2013

First, α_{opt} of the meter was searched with Algorithm 1 using $N_{length} = 29$ training data period lengths varying from 3 (45 minutes) to 100 (25 hours) each with $N_{period} = 100$ randomly chosen period locations. For each period length, α_{opt} was solved to minimize the imputation error over the period locations. This resulted in 29 α_{opt} values. To get a better estimate of the α_{opt} distribution, this process was repeated 100 times resulting in a 100×29 array of α_{opt} values. The sample distribution of α_{opt} over all missing data period lengths is depicted in Figure 18. The overall average α_{opt} was 0.1387. The distribution of the 100 α_{opt} values for each of the 29 training period lengths is visualized in Figure 19. For missing data period lengths above 8 (2 hours), the majority of

the α_{opt} values are between 0 and 0.4. For short missing data period lengths, smaller α_{opt} is preferred effectively putting more emphasis on the linear interpolation.

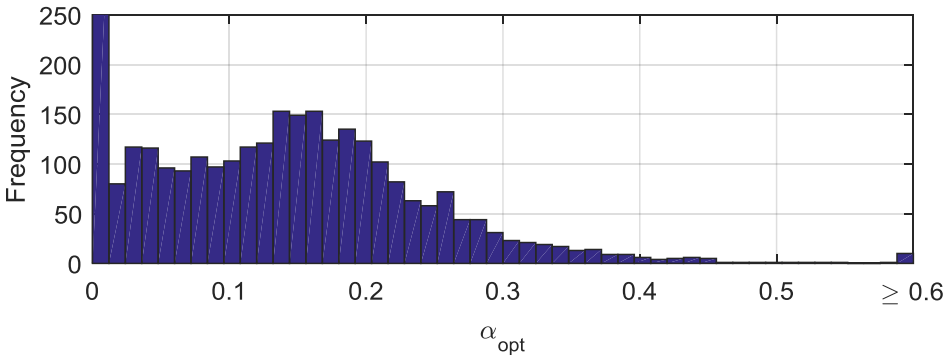

Figure 18. The distribution of the optimal weight parameter for all the training data period lengths for the analyzed Georgia Tech smart meter

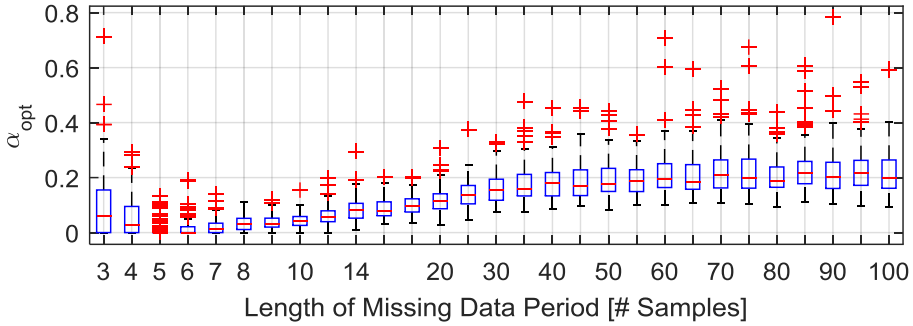

Figure 19. The boxplots of 100 optimal weight parameter values for each of the 29 different missing data period lengths for the analyzed Georgia Tech smart meter

Next, the sample mean $\alpha_{opt} = 0.1387$ was utilized to compare the performance of the OWA imputation to HA, LI, and industry best practice (BP) imputations. As discussed in Section 3.3.1, BP imputation uses LI imputed values for missing data periods shorter than 2 hours and an average of three previous days for missing data periods above 2 hours [50]. The validation was done for $N_{length} = 29$ missing data period lengths (same as used for α_{opt} training) each with 50 randomly chosen period locations. The period locations were chosen independent of the period locations used for α_{opt} training. For each period length, a mean absolute percentage error

$$MAPE = \frac{1}{N}\sum_{i=1}^{N_{sample}} \left| \hat{y}_i^{imputed} - y_i^{true} \right| \quad (17)$$

was calculated over the N_{sample} samples of the 50 missing data periods. This resulted in 29 MAPE values. To obtain a more stable estimate of the MAPE distribution, the process was repeated 100 times resulting in a 100×29 array of MAPE values. Figure 20 illustrates the distribution of the 100 MAPE values for each 29 validation data period lengths. On average, OWA outperforms HA, LI, and BP imputations for all missing data period lengths. Compared to HA and LI imputation, the advantage of OWA imputation is greater for short and long periods, respectively. For periods under 2 hours, OWA operates fairly similarly to BP but for periods over 2 hours, OWA outperforms BP imputation. Only average MAPE reduction can be expected since no imputation method is guaranteed to be effective for all missing data period lengths and characteristics.

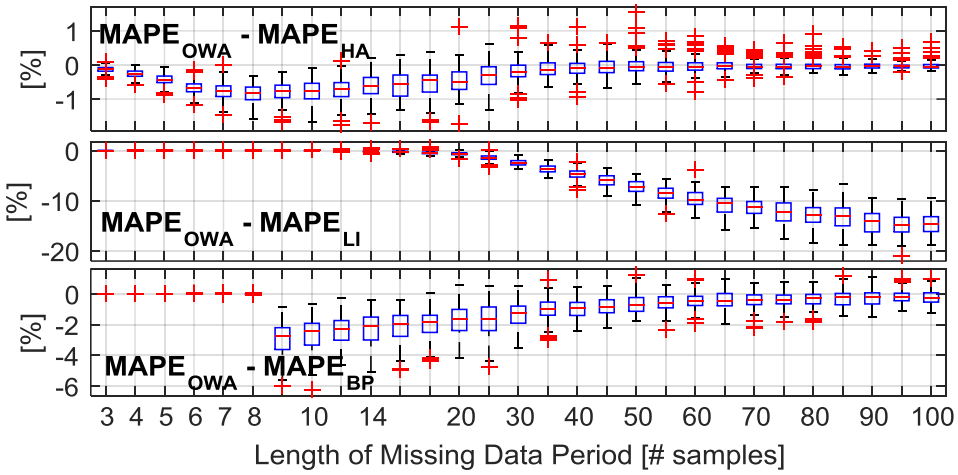

Figure 20. The boxplots of 100 MAPE differences between OWA imputation method and HA, LI, and BP imputation methods for each of the 29 different missing data period lengths for the analyzed Georgia Tech smart meter

3.3.2.2 Results for 128 Georgia Tech Smart Meters

Next, Algorithm 1 was used to search the α_{opt} for 128 Georgia Tech smart meters. For each meter, Algorithm 1 was executed with $N_{length} = 29$ training data period lengths

(same as in Section 3.3.2.1) and $N_{period} = 100$ randomly chosen missing data periods. This resulted in a 128×29 array of α_{opt} values.

The distribution of α_{opt} for the analyzed 128 Georgia Tech smart meters and all the training data period lengths is shown in Figure 21. The average α_{opt} (over all meters and all missing data period lengths) was 0.1081. The spike at $\alpha_{opt} = 0$ is caused by the bounds set on α_{opt} and spike at $\alpha_{opt} = 0.6$ represents all values $\alpha_{opt} \geq 0.6$ (including the values at the bound $\alpha_{opt} = 2.0$).

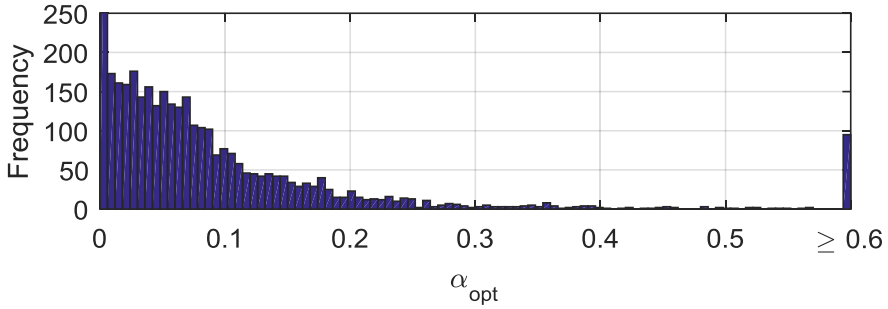

Figure 21. The distribution of the optimal weight parameter for 128 Georgia Tech meters and the 31 training data period lengths

Figure 22 visualizes the distribution of α_{opt} values for different training data period lengths and meters. As shown in the top plot of Figure 22, median α_{opt} seems to be relatively independent of the training data period lengths except for very short period lengths less than 8 (2 hours) for which α_{opt} seems to be slightly higher. The bottom plot of Figure 22 indicates that α_{opt} takes similar values for most meters but that there are also meters for which $\alpha_{opt} = 0$ or $\alpha_{opt} = 2$ for many training data period lengths. For these meters, better imputation accuracy can be achieved by solely using LI imputation or HA imputation, respectively.

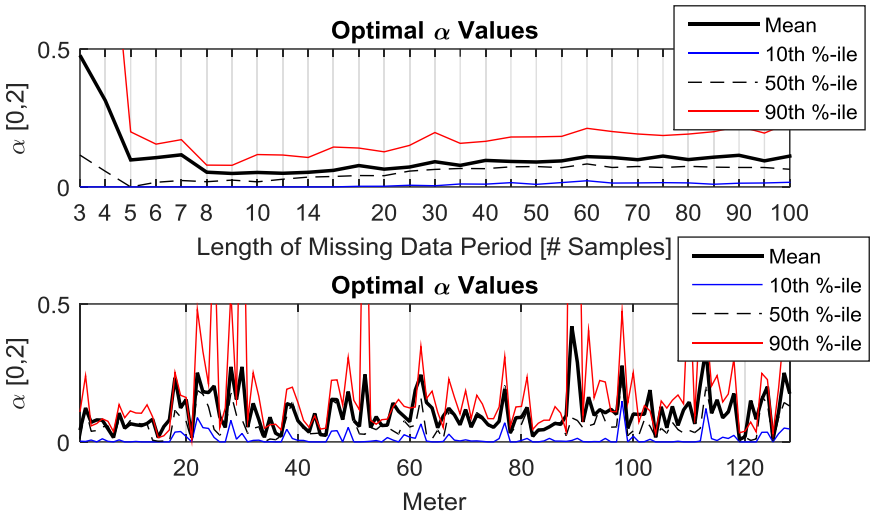

Figure 22. The percentiles of the optimal weight parameter distribution for training data period lengths (top) and for the analyzed 128 Georgia Tech smart meters (bottom)

Next, the overall average $\alpha_{opt} = 0.1081$ was utilized to compare the performance of the OWA imputation to the HA, LI, and industry best practice (BP) imputations. The validation was done for the same $N_{length} = 29$ missing data period lengths each with 100 period locations that were chosen randomly and independent of the period locations used for α_{opt} training. The distribution of MAPE differences between the OWA and the HA, LI, and BP imputations are illustrated in Figure 23. Ignoring outliers ($\geq 10\%$ and $\leq -10\%$), the average (over all missing data period lengths and meters) MAPE reductions and the respective 95% confidence intervals of the OWA approach compared to the HA, LI, and BP imputation methods were (-0.8070±0.0189)%, (-0.9831±0.0381)%, and, (-1.8592±0.0520)%, respectively.

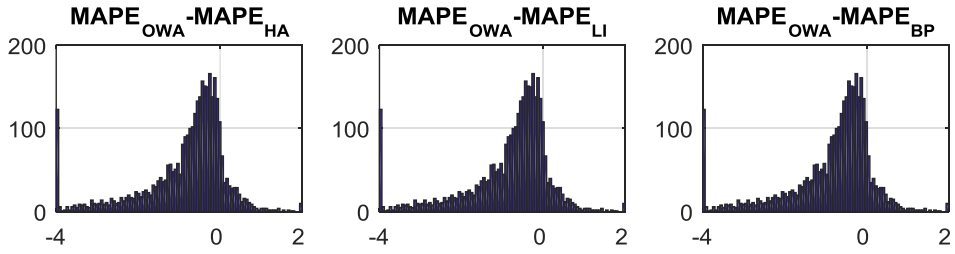

Figure 23. The histograms of MAPE differences between OWA imputation method and HA, LI, and BP imputation methods

Figure 24 shows the percentiles of the MAPE reductions for the analyzed 128 Georgia Tech smart meters. The level of MAPE reduction varies among meters but compared to HA, LI, and BP, OWA reduces the average (over the 29 missing data period lengths) MAPE values for 93.0%, 79.0%, and 93.0% of all the meters, respectively. Figure 25 visualizes the MAPE reductions for different missing data period lengths. Compared to HA, LI, and BP, OWA achieves smaller average MAPE values for medium, long, and long period lengths, respectively.

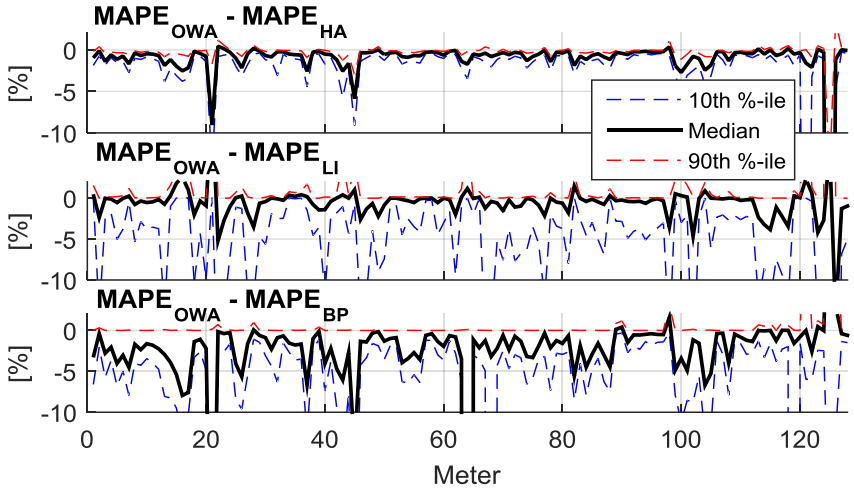

Figure 24. The percentiles of MAPE differences between OWA imputation method and HA, LI, and BP imputation methods for the 128 analyzed Georgia Tech smart meters

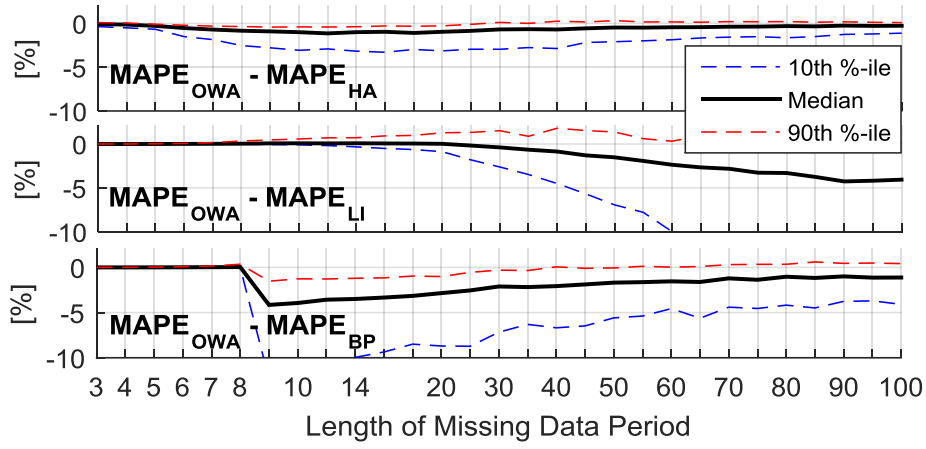

Figure 25. The percentiles of MAPE differences between OWA imputation method and HA, LI, and BP imputation methods for different missing data period lengths for the 128 analyzed Georgia Tech smart meters

3.4 Model Verification Method Using AMI Data

The smart meter voltage measurements can be compared to the simulated voltages to determine the accuracy of the model. In practice, because of modeling assumptions and measurement errors, etc., a voltage error always exists between the simulated voltages and the measured voltages. The ANSI C84.1 standard specifies steady-state voltage tolerances with a typical tolerance band of +/-5 % for service voltages (<600 V) and smaller bands for higher voltage levels [94]. In order for a distribution feeder model to be useful, the voltage error should be only a small fraction of the ANSI voltage tolerance, e.g., 10 % of the tolerance band, i.e. a voltage error of +/-0.5 %. On the other hand, the modeling and model verification accuracy is limited by smart meter (and other sensor) accuracy that is typically +/-0.2 % or +/-0.5 % for class 0.2 and 0.5 devices, respectively [39].

The accuracy of the Georgia Tech distribution system model was verified by running power flow analyses over a one-week time period by comparing the simulated voltages to the measured voltages at the loads, i.e., the meters. Figure 26 shows the meter mean bias errors (MBE)

$$\mathrm{MBE} = \frac{1}{n}\sum_{i=1}^{n}\frac{(V_{sim} - V_{meas})}{V_{meas}} \tag{18}$$

between the simulated and measured voltages. Positive values indicate that the simulated voltages were higher than the measured voltages, which was likely caused by circuit impedances that were too low. The general accuracy of any metered point in the model can be given as the standard deviation of the voltage simulation error for that meter, each of which are shown as the error bars Figure 26. Figure 27 presents the voltage simulation error variances on the feeder voltage profile pointing out an acceptable range of voltage simulation accuracy.

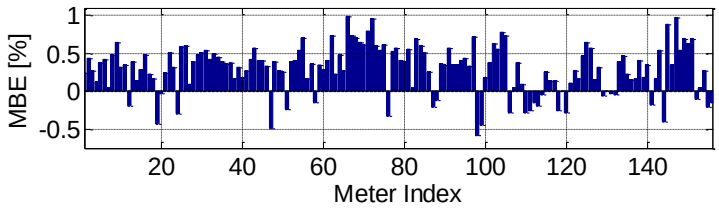

Figure 26. The mean bias error of simulated load voltages compared to measured load voltages from April 27th, 2014 through May 3rd, 2014

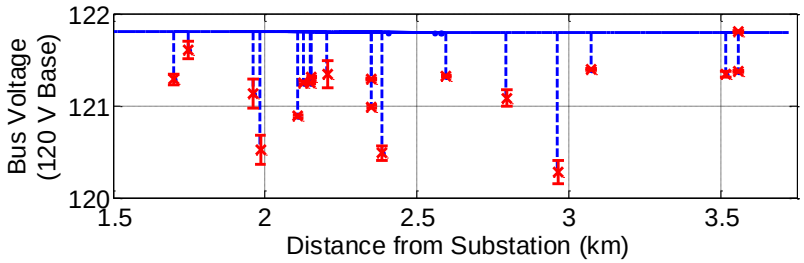

Figure 27. A feeder voltage profile with red error bars for the standard deviations of the simulation errors of the day

These voltage verification methods illustrate that the Georgia Tech distribution system model reasonably represents the physical system. Either incorrect model assumptions or major errors in the voltage measurements would be very apparent in the results. While this verifies the general model configuration, the differences between the measured and simulated voltages can be decreased by estimating the secondary circuit model parameters and topologies with the approaches discussed in Chapters 4-6.

3.5 Discussion

The presented data validation methods for smart meter applications are effective in identifying many typical smart meter data accuracy issues. However, many issues, such as wrong meter CT ratio, cannot be detected without having either detailed information of the measured quantity or other redundant measurements. In the future, data validation may become a part of distribution system state estimation. However, before distribution state estimators become a norm, the methods presented in this chapter play an important role in leveraging the smart meter measurements.

The proposed data imputation method has a superior average accuracy compared to linear interpolation, historic averages, and an industry best practice mix of both. The proposed method results in imputed samples that have a continuous profile with respect to the surrounding measurements, which is a desired feature for time series applications including quasi-static time series power flow analysis [95]. The weight parameter of the proposed approach can be trained offline after which missing data can be imputed with only limited computational and data requirements. Additionally, only one model parameter needs to be stored for each meter. It is important to note that although the proposed imputation method performs better on average, it may have a lower accuracy on individual time intervals.

This chapter has shown the data validation and imputation for smart meter active power data. However, many of the proposed ideas can also be utilized for smart meter reactive power and voltage data as well as for the data from other modern distribution sensors such as PV micro-inverters. Next to the approaches shown in this chapter, PV irradiance data validation and imputation may also leverage clear sky models [96].

There are several interesting directions for future work on data imputation. First, the proposed data imputation method could be improved to better handle short missing data periods. This could be achieved by, e.g., utilizing an offset parameter for each weight function forcing all the weight for the LI imputation for short time intervals. The relationship between the proposed imputation method and conventional statistical prediction model averaging methods such as stacking could also be investigated. Moreover, the proposed imputation method could be extended to leverage the energy readings over the missing data period when they are available.

CHAPTER 4. DISTRIBUTION SECONDARY CIRCUIT PARAMETER ESTIMATION WITH AMI DATASETS

This chapter presents distribution system parameter estimation (DSPE) problem with AMI measurements at all loads and proposes a linear regression parameter estimation method (LRPE) that is applied to modeling of secondary circuits. This chapter has the following structure. First, section 4.1 defines the DSPE problem. Then, section 4.2 proposes a linear regression based method for estimating the (positive sequence series) impedance parameters of circuit subsection consisting of either a single transformer/line or a number of parallel transformers/lines. Next, section 4.3 presents the LRPE algorithm that generalizes the parameter estimation of the simple circuit subsections to entire radial secondary circuit parameter estimation. Section 4.4 discusses measurement sample selection for parameter estimation and section 4.5 compares alternative linear regression models and proposes an optimal adaptive linear regression model approach for parameter estimation. Sections 4.6 and 4.7 demonstrate the LRPE method on a three-phase test circuit and on one of the Georgia Tech feeders. Most of the work covered in this chapter has been published in [97]–[99].

4.1 Problem Definition

The objective of DSPE is to find the most likely values of resistance (R) and reactance (X) parameters of a distribution circuit. In this chapter we address this problem for the secondary portion of the circuit, as illustrated in Figure 28. The method assumes that historical voltage (V), active power (P), and reactive power (Q) measurements shown in blue in the figure are available at all the leaf nodes of the secondary circuit tree. This assumption holds if the corresponding customers have AMI data. In order to estimate the service transformer parameters, the method requires service-transformer medium-voltage

values. In some cases those measurements exist. If not, the service transformer values can be obtained from simulations of the primary circuit.

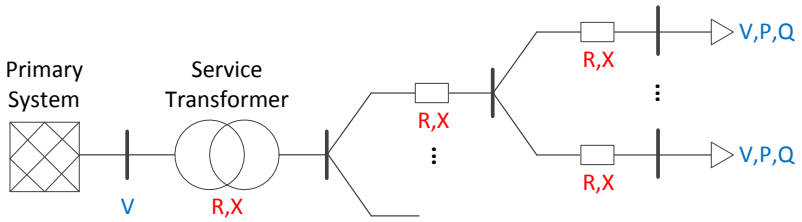

Figure 28. Secondary circuit tree for parameter estimation

The proposed method relies on the following four assumptions.

1. The secondary circuit topology is assumed to be known. Chapter 6 describes a method that can be used if the topology is unknown. The secondary circuit is assumed to be radial (i.e. a tree) like most real secondary circuits [100], [101].
2. The active and reactive power (or current and power factor) and the voltage measurements are assumed to be available at all leaf nodes of the tree. In practice, this assumption is valid for secondary circuits with AMI since most secondary circuit tree leaf nodes have either a load and/or a distributed generation (DG) unit with the respective measurements. Handling cases where some meters report no voltage measurements is discussed separately in section 5.1.
3. The secondary circuit is assumed to be either balanced 3-phase or single-phase. This assumptions is often invalid since in practice many distribution system secondary circuits are split-phase, i.e., a single-phase where a center-tapped transformer connects to a triplex cable with both 120V and 240V service to the loads. Although it is possible to model the split-phase secondary circuits in detail [102], parameter estimation is limited by the available measurement data, which typically consists of the customer total power and/or current as well as voltage measurement across the 120V (or the 240V) connection. As long as the power, current and voltage measurements for both the 120V and 240V loads are not

included in the MDMS, it may be desirable to model split-phase secondary circuits with single-phase transformers, lines, and loads. Using this modeling approach, typical measurement meter data can be readily utilized to estimate the secondary circuit transformer and line parameters utilizing the LRPE method introduced below.

4.2 Linear Regression Branch Series Impedance Parameter Estimation Method

The proposed linear regression branch parameter estimation method is based on the well-known (see e.g. [97], [100], [103]) linear approximation of voltage drop magnitude over a series impedance shown in Figure 29 on the right

$$V_{drop} = V_1 - V_2 \approx (RP + XQ)/V_2 = RI_R + XI_X, \tag{19}$$

where V_1 and V_2 are voltage magnitudes, R and X are the series resistance and reactance between two buses (positive sequence for balanced 3-phase branches and phase impedance for 1-phase branches). The current resistive and reactive components are given with

$$I_R = P/V_2 = I(PF) \tag{20}$$

and

$$I_X = Q/V_2 = I\sqrt{1-(PF)^2}, \tag{21}$$

where (PF) is the power factor. For transformers, all values must be referred to the same voltage level. In 3-phase systems, line-line voltages and 3-phase powers are used whereas in 1-phase systems, line-to-neutral voltages are utilized.

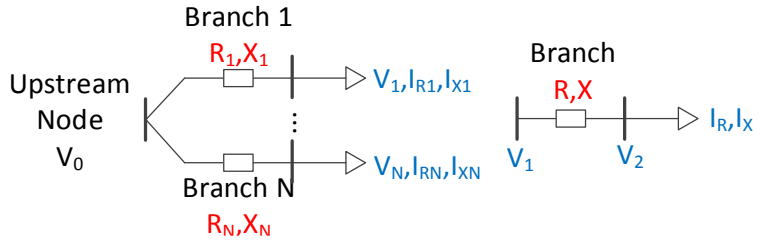

Figure 29. Secondary circuit section with N parallel branches (left) and no parallel branches (right)

4.2.1 Single Branch Parameter Estimation

This section presents a linear regression branch (positive sequence series impedance) parameter estimation method for a circuit section shown in Figure 29 on the right. The goal of the method is to find the parameters R and X (shown in red in Figure 29 on the right) that provide the best fit of the M measurement samples of V_1, V_2, P, and Q (shown in blue in in Figure 29 on the right) to the linear model

$$\Delta \boldsymbol{V} = \boldsymbol{V}_1 - \boldsymbol{V}_2 = R\boldsymbol{I}_R + X\boldsymbol{I}_X + \boldsymbol{\epsilon}, \tag{22}$$

where $\boldsymbol{\epsilon} \in \mathbb{R}^M$ captures the model and measurement error and all the bold letters indicate vectors (or matrices) through time. Denoting the response vector $\mathbf{y} \in \mathbb{R}^M$, the design matrix $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{M\times 2}$, and the unknown parameter vector $\boldsymbol{\beta} \in \mathbb{R}^2$

$$\boldsymbol{y} = \boldsymbol{V}_1 - \boldsymbol{V}_2, \tag{23}$$

$$\boldsymbol{\mathcal{X}} = [\boldsymbol{I}_R \quad \boldsymbol{I}_X], \tag{24}$$

and

$$\boldsymbol{\beta} = [R \quad X]^{\mathrm{T}} \tag{25}$$

respectively, gives the linear model

$$\boldsymbol{y} = \boldsymbol{\mathcal{X}}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{26}$$

An estimate for the unknown parameters, $\widehat{\boldsymbol{\beta}}$, can be obtained by, e.g., minimizing the p-norm of the model residuals over the measurement samples

$$\widehat{\boldsymbol{\beta}} = \underset{R,X}{\operatorname{argmin}} \|\boldsymbol{y} - \boldsymbol{\mathcal{X}}\boldsymbol{\beta}\|_p. \tag{27}$$

If $p = 1$, (27) becomes a linear programming problem whose solution is referred to as the least absolute value (LAV) estimator. LAV has the advantage of being insensitive to outliers but has no closed-form solution and requires solving a linear programming problem. With $p = 2$, (27) is a linear unconstrained least squares problem whose solution is the ordinary least squares (OLS) estimator given as the solution of the normal equations

$$\boldsymbol{\mathcal{X}}^{\mathrm{T}}\boldsymbol{\mathcal{X}}\widehat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{\mathcal{X}}^{\mathrm{T}}\boldsymbol{y}. \tag{28}$$

Under a handful of conditions, the OLS estimator $\widehat{\boldsymbol{\beta}}_{OLS}$ has very attractive properties. First, $\widehat{\boldsymbol{\beta}}_{OLS}$ is a consistent estimator, i.e., as the sample size grows, the estimated parameters approach the true parameters. OLS is also unbiased (i.e. $\mathbb{E}\left[\widehat{\boldsymbol{\beta}}_{OLS}\right] = \boldsymbol{\beta}_{true}$) and has the minimum variance among all unbiased estimators [104]. Under the further assumption of independent, normally distributed errors, OLS is equal to the maximum likelihood estimator [104]. One of the disadvantages of OLS is that the results tend to be sensitive to outliers, which is why an effective outlier detection and removal is essential. It should be noted that in order to solve (28) for $\widehat{\boldsymbol{\beta}}_{OLS}$, the design matrix $\boldsymbol{\mathcal{X}}$ must have a full column rank, i.e., the predictors (columns) of $\boldsymbol{\mathcal{X}}$ must be linearly independent. In practice, this means that the net loads of all leaf nodes must differ from each other (no net loads equal to zero).

Sometimes it is desirable to set bounds on the impedance parameter estimates. This can be done by utilizing linearly constrained least squares estimator (CLS) given by

$$\begin{gathered}\widehat{\boldsymbol{\beta}}_{CLS} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\mathcal{X}}^{\mathrm{T}}\boldsymbol{\mathcal{X}}\boldsymbol{\beta} \\ \text{subject to } \boldsymbol{C\beta} \leq \boldsymbol{d},\end{gathered} \tag{29}$$

where the constraint $\boldsymbol{C\beta} \leq \boldsymbol{d}$ describes the parameter bounds. This quadratic (convex) programming problem has no closed-form solution but can be effectively solved to a global optimum with any open-source or commercial solver. The global optimum is unique provided that the Gramian matrix $\boldsymbol{\mathcal{X}}^{\mathrm{T}}\boldsymbol{\mathcal{X}}$ is positive definite, which takes place if and only if the predictors of $\boldsymbol{\mathcal{X}}$ are linearly independent. In this dissertation the CLS estimator is only used when the OLS estimator results in non-physical, e.g., negative or too high, parameter values.

4.2.2 Parallel Branch Parameter Estimation

This section presents a linear regression (positive sequence series impedance) parameter estimation method for the circuit section with $N \in \{2,3,\dots\}$ parallel branches shown in Figure 29 on the left. In this circuit section, the upstream voltages V_0 are unknown. Similarly to the single branch parameter estimation, the goal is to find the most likely parameters $R_i, X_i, i \in \{1,\dots,N\}$ (shown in red in Figure 29 on the left). This parameters are selected to find the best fit (in the least-squares sense) of the M measurement samples of $V_i, P_i, Q_i, i \in \{1,\dots,N\}$ to the linear model

$$\boldsymbol{y} = \boldsymbol{\mathcal{X}}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{30}$$

where $\boldsymbol{\epsilon} \in \mathbb{R}^M$ represents the measurement and model error, $\boldsymbol{\beta} \in \mathbb{R}^{(M+2N)}$ is the parameter vector given by

$$\boldsymbol{\beta} = \left[V_{0,1}, \dots, V_{0,M}, R_1, X_1, \dots, R_N, X_N\right]^{\mathrm{T}}, \tag{31}$$

and $\boldsymbol{y} \in \mathbb{R}^{MN}$ is the response vector is given by

$$\boldsymbol{y} = \left[V_{1,1}, \dots, V_{1,M}, \dots, V_{N,1}, \dots, V_{N,M}\right]^{\mathrm{T}}. \tag{32}$$

Finally, the design matrix $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{(MN)\times(M+2N)}$ is given by

$$\boldsymbol{\mathcal{X}} = \begin{bmatrix} \mathbf{I} & \boldsymbol{J}_1 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{I} & \mathbf{0} & \cdots & \boldsymbol{J}_N \end{bmatrix}, \tag{33}$$

where $\mathbf{I} \in \mathbb{R}^{M\times M}$ are identity matrices, the submatrices $\boldsymbol{J}_i = [-\boldsymbol{I}_{R,i} \quad -\boldsymbol{I}_{X,i}] \in \mathbb{R}^{M\times 2}, i \in \{1,\dots,N\}$ consist of the branch current measurements, and the zero submatrices $\mathbf{0}$ have suitable sizes. This formulation has MN equations and $(M+2N)$ unknowns (excluding the error terms). In practice, a large sample size ($M \gg N$) is used resulting in a large but extremely sparse $\boldsymbol{\mathcal{X}}$. Therefore, sparse matrix methods should be utilized when operating with the $\boldsymbol{\mathcal{X}}$ in (33). In order to obtain the OLS estimate of the unknown parameters, $\widehat{\boldsymbol{\beta}}$, the system of normal equations in (28) has to be solved, where $\boldsymbol{\mathcal{X}}^{\mathrm{T}}\boldsymbol{y} \in \mathbb{R}^{M\times 1}$ is a full vector

(since $\boldsymbol{y}$ is full). The design matrix $\boldsymbol{\mathcal{X}}$ defined in (33) results in a Hermitian positive-definite arrowhead block matrix

$$\boldsymbol{\mathcal{X}}^{\boldsymbol{T}}\boldsymbol{\mathcal{X}} = \begin{bmatrix} \mathrm{N}\times\mathrm{I} & \boldsymbol{J}_1 & \cdots & \boldsymbol{J}_N \\ \boldsymbol{J}_1^T & \boldsymbol{J}_1^T\boldsymbol{J}_1 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{J}_N^T & \mathbf{0} & \cdots & \boldsymbol{J}_N^T\boldsymbol{J}_N \end{bmatrix}, \tag{34}$$

whose Cholesky decomposition $\boldsymbol{L}$ ($\boldsymbol{\mathcal{X}}^{\boldsymbol{T}}\boldsymbol{\mathcal{X}} = \boldsymbol{L}\boldsymbol{L}^{\boldsymbol{T}}$, where $\boldsymbol{L}$ is a lower-triangular matrix with real and positive diagonal entries) remains extremely sparse [105]. The OLS estimator of the unknown parameters $\widehat{\boldsymbol{\beta}}_{OLS}$ can be efficiently obtained by solving the two triangular systems by forward and backward substitution $\boldsymbol{L}\boldsymbol{Y} = \boldsymbol{y}$ and $\boldsymbol{L}\widehat{\boldsymbol{\beta}} = \boldsymbol{Y}$, respectively. Matlab does this be default for a linear system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, where $\boldsymbol{A}$ is a real symmetric matrix.

It should be noted that the column ordering of $\boldsymbol{\mathcal{X}}$ has a significant impact on the computational complexity of solving the normal equations. For example, if the columns of $\boldsymbol{\mathcal{X}}$ are ordered

$$\boldsymbol{\mathcal{X}} = \begin{bmatrix} \boldsymbol{J}_1 & \cdots & \mathbf{0} & \mathrm{I} \\ \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \cdots & \boldsymbol{J}_N & \mathrm{I} \end{bmatrix}, \tag{35}$$

then the Cholesky decomposition of the matrix $\boldsymbol{\mathcal{X}}^{\boldsymbol{T}}\boldsymbol{\mathcal{X}}$ results in a very large number of fill-ins and therefore, the normal equations are much more demanding to solve. To conclude, the columns of $\boldsymbol{\mathcal{X}}$ should be ordered as shown in (33). Figure 30 illustrates the sparsity of the matrices $\boldsymbol{\mathcal{X}}$, $\boldsymbol{\mathcal{X}}^{\boldsymbol{T}}\boldsymbol{\mathcal{X}}$, and the Cholesky decomposition of $\boldsymbol{\mathcal{X}}^{\boldsymbol{T}}\boldsymbol{\mathcal{X}}$ both for the column order of $\boldsymbol{\mathcal{X}}$ shown in (33) and the alternative column order of $\boldsymbol{\mathcal{X}}$ shown in (35). The column order in (35) results 400 times more fill-ins than the order in (33).

Once the impedances, currents, and downstream node voltages of the N parallel branches are known, the voltages of the upstream node of the branches can be estimated as an average of the upstream node voltage estimates of the N branches with

$$\boldsymbol{V}_0 = \frac{1}{N}\sum_{i=1}^{N}\left\|\boldsymbol{V}_i + (R_i + jX_i)(\boldsymbol{I}_{R,i} + j\boldsymbol{I}_{X,i})\right\|. \tag{36}$$

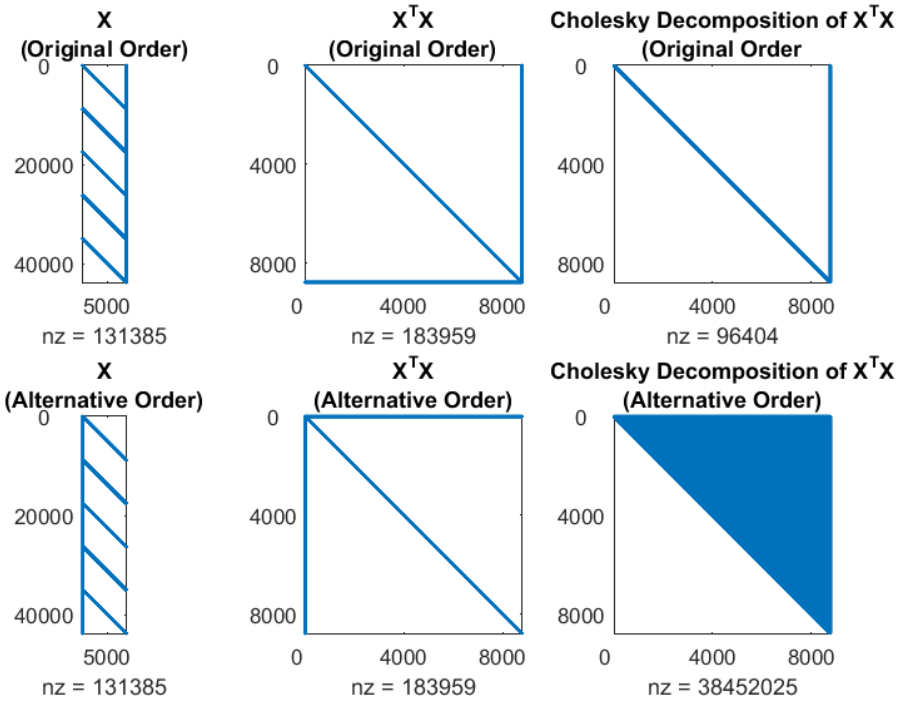

Figure 30. The influence of column order for the sparsity of the parallel branch linear regression matrices

4.3 Algorithm

This section presents the LRPE algorithm. The algorithm processes one secondary circuit tree at a time, hierarchically proceeding from the tree leaf nodes towards the tree root node. At a given iteration the algorithm estimates the branch impedances for a subsection of the secondary circuit using linear regression as follows. First, the algorithm searches for a circuit subsection with one of the two types discussed in section 4.2, whose parameters are not estimated yet. The selected subsection can be either A) a single branch that has known (measured or estimated at previous iteration) upstream and downstream node voltages and downstream node currents shown in blue in Figure 29 on the right, or B) a set of parallel branches with known (measured or estimated at previous iteration) downstream node voltages and currents shown in blue in Figure 29 on the left. Once a

suitable circuit subsection has been identified, the algorithm first estimates the subsection branch impedance parameters with the approaches shown in section 4.2 and then in case of the parallel branch case, estimates the upstream node voltages with (36) using the measurements and the estimated branch parameters. These steps are listed in detail in Algorithm 2.

Algorithm 2. LRPE algorithm

Input: A list of all secondary circuit branches, $\mathcal{L}$, with fields upstream and downstream node, number of parallel branches, branch current measurements $\boldsymbol{I}_R, \boldsymbol{I}_X$, branch node voltage measurements $\boldsymbol{V}$
Output: Secondary circuit branch estimation results including the estimated parameters R_{est}, X_{est} and their p-values R_{pval}, X_{pval}, and the regression model R^2- and MSE-values
Initialization: Set the list of active branches, ℓ, empty.
1. IF $\mathcal{L}$ is empty, **STOP**.
2. IF $\mathcal{L}$ has only one branch i with both upstream and downstream voltage measurements or estimates, set $\ell = i$ and remove ℓ from $\mathcal{L}$.
ELSEIF $\mathcal{L}$ has a branch i whose $N-1$ parallel branches $j_1, \dots, j_{N-1}$ have downstream node voltage measurements or estimates. Set $\ell = \{i, j_1, \dots, j_{N-1}\}$.
ELSE Print warning that the circuit does not have sufficient measurement points and **STOP**.
ENDIF
3. IF ℓ has only one branch, estimate the impedance parameters of the branch in ℓ with the single-branch regression formulation (23)-(26) and go to **1.**
ELSE Estimate the impedance parameters of the branches in ℓ with the parallel branch regression formulation (30)-(33), estimate the voltages of the upstream node of the N parallel branches with (36), and go to **1.**
ENDIF

4.4 Measurement Sample Selection for Parameter Estimation

Since the parameter estimation algorithm is run off-line utilizing historical data, it is possible to selectively pick a subset of the available measurement samples. First, when any of the necessary meters has missing or bad data, all simultaneous measurement samples

should be ignored from the linear regression parameter estimation. Second, bad data can be detected with conventional approaches such as checking for unrealistically high or low values based on historical data [50].

The linear regression parameter estimation presented above, utilizes the voltage drop magnitude approximation that is well-known to be quite accurate for typical P, Q, R, and X values [103]. The largest error occurs under heavy load (current) and leading power factor [103]. The relative linearization error with respect to P and Q for line with an X/R=1 is shown in the top right plot of Figure 31. With typical P and Q combinations, the error is below 1-2%, but it can be significantly higher with either 1) large positive P and small negative Q or 2) large negative P and small positive Q. While Case 1 is very untypical in distribution secondary systems where most loads are inductive, Case 2 can occur in secondary systems when a large injection from distributed generation at unity power factor leads to reverse active power flow while the inductive loads consume VArs.

The relative linearization error with respect to R and X for a 50kVA load at power factors $(\mathrm{PF}) = \{0.9, 0.95, 0.98, 1.0\}$ are shown in Figure 32. Typical distribution system secondary circuit X/R ratio is in the range of 1 to 2, resulting in a linearization error below 2%. However, for circuits with high X/R ratio, the accuracy can be considerably worse.

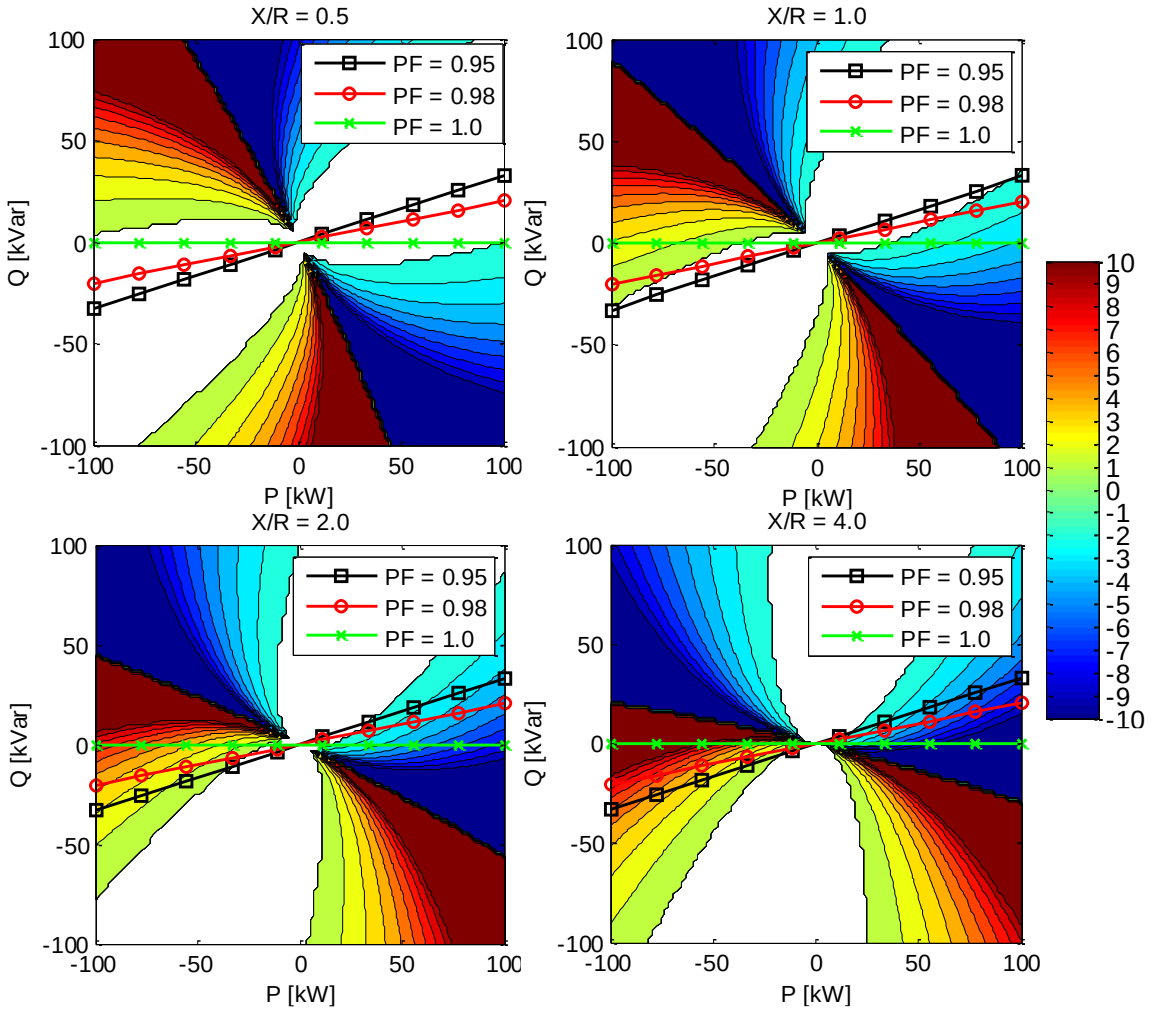

Figure 31. Voltage drop linearization error [%] for a range of P and Q with X/R={0.5,1.0,2.0,4.0}, white areas have error ≤1%, error magnitudes ≥10% are set to 10%

Provided that sufficient data is available, the figures suggest that the following data be filtered before parameter estimation: 1) Measurement samples that have both high P demand and power factor above 0.95 or 0.98 and 2) Measurement samples that have both high P generation (reverse power flow) and Q consumption. Filtering data has the disadvantage of reducing the number of available measurement samples, which can reduce the parameter estimation accuracy. Therefore, the filtering should only be considered for samples that have a considerable negative impact on the estimation accuracy. Whether the sample filtering is advantageous or not may depend on the characteristics of the load data

at hand. The next section presents and compares alternative linear regression models that can be used to partially compensate the error in the linearized voltage drop equation and to handle error in practical measurement data sets.

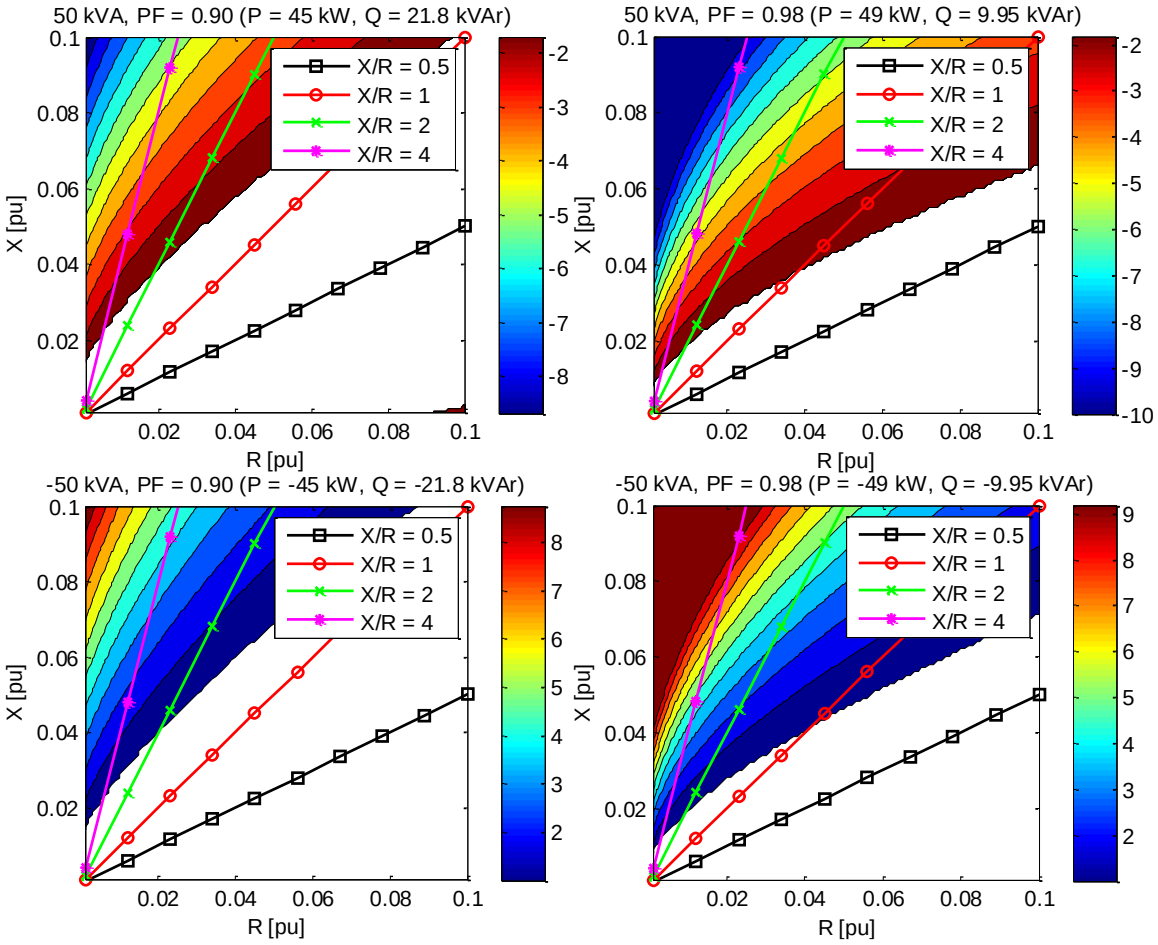

Figure 32. Voltage drop linearization error [%] for a range of R and X with S = ±50kVA, (PF)={0.95,0.98}, white areas have error ≤1%, error magnitudes ≥10% are set to 10%

4.5 Optimal Linear Regression Model

The linear regression models (23)-(26) and (30)-(33) are based on the linearized voltage drop equation (22) and thus, the predictors and unknown parameters have direct physical meanings in both formulations. However, generalized linear regression allows models with higher order terms, cross-couplings, or any other functions of the predictor variables I_R and I_X. Unlike R and X in (25) and (31), the coefficients of other terms do not

have a direct physical meaning, but including them in the regression models may better capture the intrinsic nonlinear relationship between the response variable and the predictor variables thus, leading to better estimates for R and X. This section analyzes the accuracy of different regression models (i.e. different combinations of predictor variables) by estimating all the secondary circuit parameters in a test circuit and calculating the average absolute error of the estimated R and X parameters. First, a 3-phase test circuit, which was implemented to test and validate parameter estimation work, is introduced. Then, sections different linear regression models are analyzed without and with measurement error. The section concludes by proposing an optimal adaptive approach for regression model selection.

4.5.1 Three-Phase 66-Node Test Circuit Overview

A 66-node test system was created to demonstrate the parameter estimation performance with various secondary circuit topologies and component parameters. The circuit has a 3-phase 12kV L-L backbone feeder, ten 3-phase 240V L-L secondary circuits each with a different topology listed in Table 3, and 36 loads in the secondary circuits. The total capacity of the ten service transformers is 600kVA. Each load was assigned a measured active power profile from [106] and a random power factor profile sampled from the uniform distribution PF∼Unif[0.9,1.0]. Typical primary and secondary line and transformer parameters were used. The circuit voltage profile at a measurement sample is shown in Figure 33 on the left. The circuit topology is illustrated in Figure 33 on the right contouring showing per-unit voltages and line widths showing current magnitudes.

Table 3. Secondary circuits of the 66-node three-phase test circuit

Secondary Circuit Number (Order from the Substation)	Load Connections
1	5 loads connected to the transformer
2	1 large load connected to a pedestal
3	5 loads connected in series on a service line (without separate service drops)
4	5 loads connected to a pedestal
5	2 loads connected to the transformer and 2 loads connected to a pedestal
6	2 separate pedestals each with two loads
7	2 pedestals in series each with 2 loads
8	2 pedestals in series: first with one load, second with 3 loads
9	1 pedestal with two loads
10	1 load connected to the transformer, 1 pedestal with 1 load

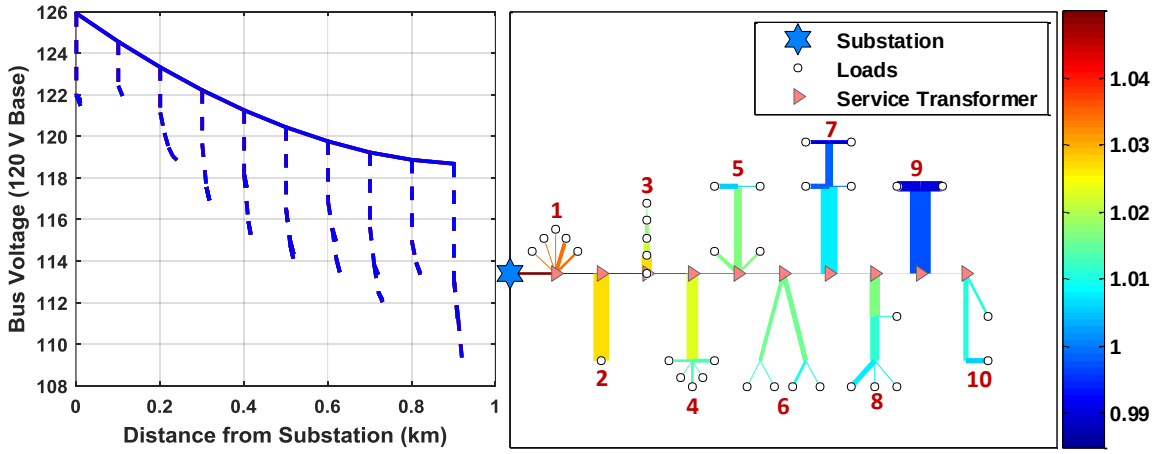

Figure 33. Three-phase test circuit voltage profile (left) and circuit line diagram contouring showing per-unit voltages and line widths showing current magnitudes (right)

The service transformer MV side voltages were assumed to be accurately simulated (accurate primary system model) from a time series load flow. Moreover, in this chapter, the secondary network topologies were also assumed to be known, and the hourly active power, reactive power, and voltage measurements of all loads are available from the AMI.

Before estimating the secondary circuit parameters with Algorithm 2, it is necessary to merge all the series branches that have no measurements in between, e.g., the service transformer and the service drop of secondary circuit no. 2 (Figure 33). Figure 34 shows

the merged secondary circuit tree topologies that are processed based on the OpenDSS circuit model. The node names and the node upstream branch names are shown on the lower left and upper right sides of the nodes, respectively. The transformer medium voltage and low voltage side nodes are abbreviated with “HV” and “LV’, respectively. The circuits where the transformer is merged with its downstream branch do not have a node with “LV”. Branches that include a transformer have “T” and branches that include a line have “L” and in their label.

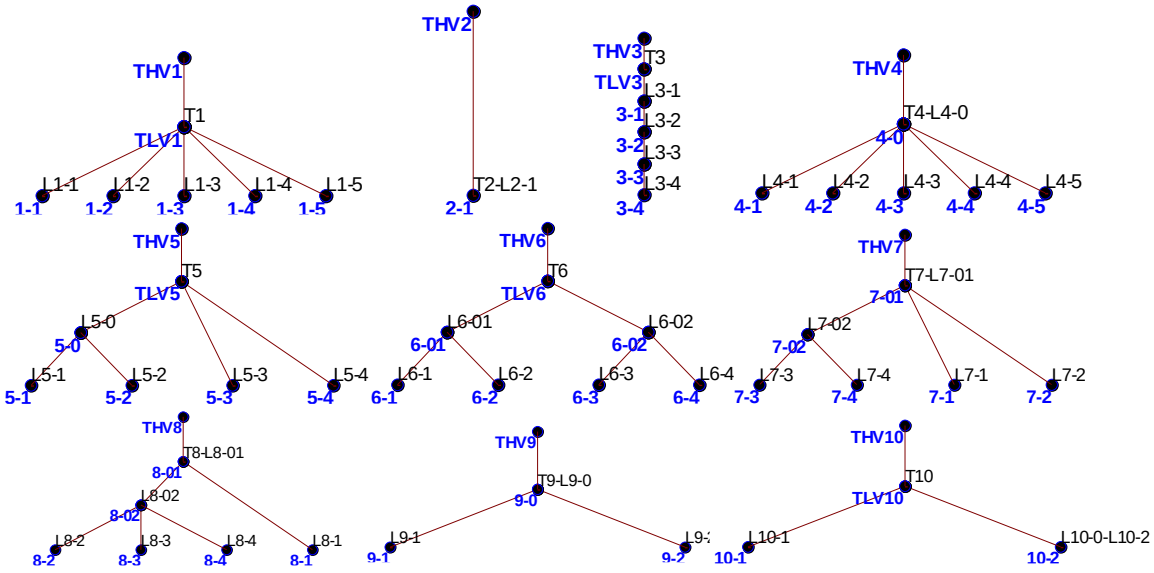

Figure 34. Merged secondary circuit trees with bus names shown in blue and bus upstream branch names shown in black

4.5.2 Regression Model Comparison without Measurement Error

First, the accuracy of different regression models was analyzed without measurement error. All the secondary circuit parameters in the 66-node test circuit were estimated and the average absolute error of the estimated R and X parameters was calculated. Table 4 compares the errors of the parameters estimated with different regression models using 8759 measurement samples (one year of hourly measurement samples). The meters are assumed to be perfectly accurate and able to record voltage and

current without any measurement error. The results are listed in the order from the simplest regression model to the most complicated. The best results (in terms of the absolute average $|R_{err}| + |X_{err}|$) are obtained with regression model $\Delta V \sim I_R + I_X + I_R \times I_X + I_R^2 + I_X^2 - 1$ (linear, quadratic and cross-coupling terms of the current components but no intercept or power factor terms). In all cases, regression models with intercept performed slightly worse than the respective models without the intercept.

Table 4. Relative parameter estimation errors for different linear regression models without measurement error

Included Predictor Variables (All Models Include I_R and I_X)					Avg. Abs. R_{err} [%]			Avg. Abs. X_{err} [%]			Avg. Abs. $\lvert R_{err}\rvert + \lvert X_{err}\rvert$ [%]	Max. R_{err} [%]	Max. X_{err} [%]	Model Order (Best to Worst)
Inter-cept	$I_R \times I_X$	I_R^2	I_X^2	(PF)	Lines	Trafos	All	Lines	Trafos	All				
					0.517	7.833	2.010	0.349	1.471	0.578	2.588	18.548	2.671	11
X					0.588	12.712	3.063	0.349	1.553	0.594	3.657	29.941	2.739	12
		X			0.151	1.502	0.426	0.350	1.593	0.604	1.030	2.328	2.781	8
X		X			0.169	1.584	0.458	0.350	1.592	0.604	1.062	3.969	2.781	9
		X	X		0.088	0.522	0.176	0.213	0.603	0.293	0.469	0.839	1.283	5
X		X	X		0.106	0.861	0.260	0.213	0.607	0.294	0.554	2.516	1.283	6
	X	X	X		0.035	0.165	0.062	0.026	0.137	0.049	0.111	0.570	0.408	1
X	X	X	X		0.056	0.618	0.171	0.026	0.144	0.050	0.221	1.994	0.413	2
		X	X	X	0.184	0.900	0.330	0.476	0.553	0.492	0.822	2.654	1.948	7
X		X	X	X	0.209	1.755	0.524	0.508	1.867	0.785	1.310	4.177	3.313	10
	X	X	X	X	0.066	0.657	0.187	0.060	0.168	0.082	0.269	2.168	0.411	3
X	X	X	X	X	0.074	0.980	0.259	0.085	0.483	0.167	0.425	2.965	1.021	4

The errors of parameters R and X estimated with regression model $\Delta V \sim I_R + I_X - 1$ and $\Delta V \sim I_R + I_X + I_R \times I_X + I_R^2 + I_X^2 - 1$ are shown in Figure 35 and Figure 36, respectively. Each bar represents a low-voltage branch in the 66-node test circuit. Branch names that start with L are lines, and branch names that start with T include service transformers. The errors of the estimated impedance magnitude Z and X/R-ratio for model $\Delta V \sim I_R + I_X - 1$ and $\Delta V \sim I_R + I_X + I_R \times I_X + I_R^2 + I_X^2 - 1$ are shown in Figure 37 and Figure 38, respectively. Without measurement error, regression model $\Delta V \sim I_R + I_X + I_R \times$

$I_X + I_R^2 + I_X^2 - 1$ estimates all the parameters with a very high accuracy. Regression model $\Delta V \sim I_R + I_X - 1$ estimates the line parameters with a relatively good accuracy but does poorly especially in estimating the service transformer resistances and X/R-ratios. The transformer R parameters are clearly over-estimated while the transformer X parameters are clearly under-estimated. An explanation for this is the linearized voltage drop approximation illustrated in Figure 32, where the higher the X/R-ratios and the impedance magnitudes are, the more the linearized voltage drop equation underestimates the voltage drop. As a result, the transformer resistances will be over-estimated and the reactances under-estimated in the linear regression parameter estimation. This is the direction where the voltage drop approximation error reduces the fastest.

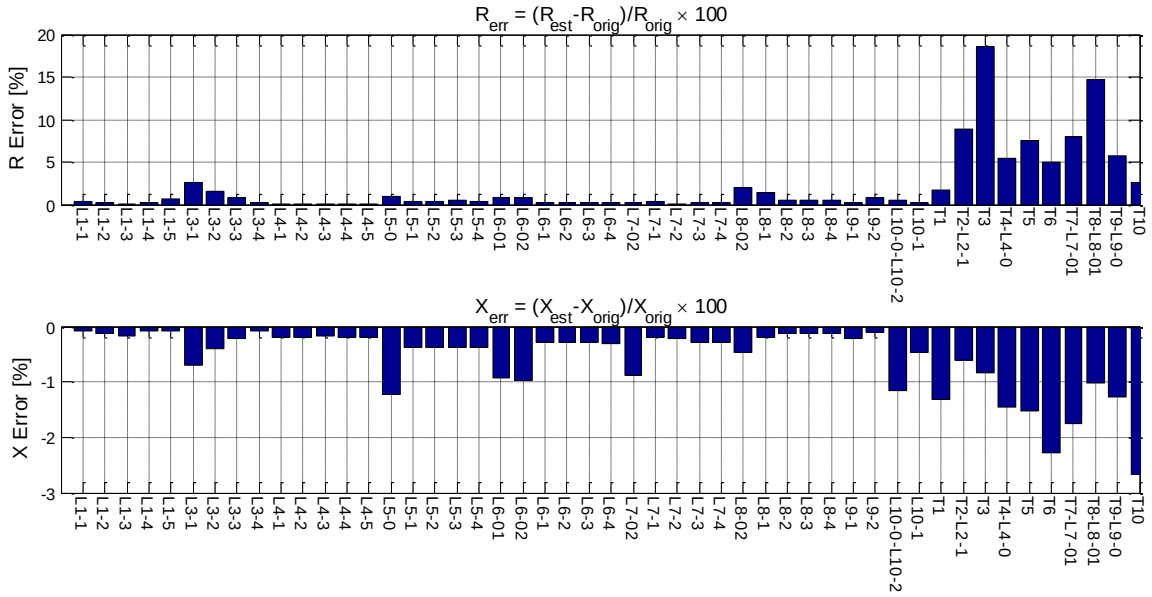

Figure 35. Relative errors of estimated R and X with regression model $\Delta V \sim I_R + I_X - 1$ without measurement error

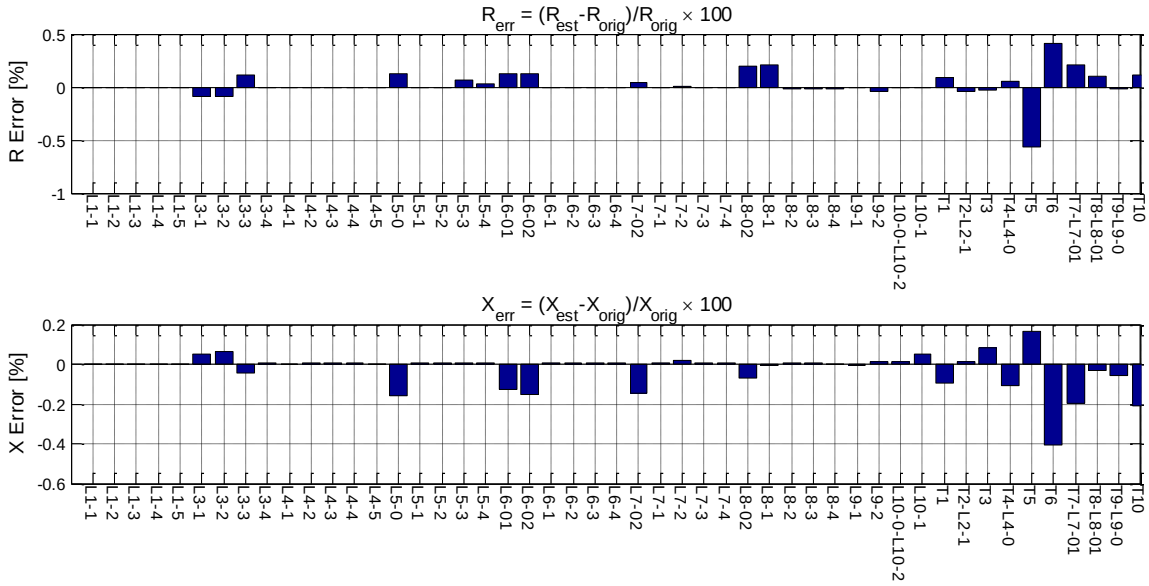

Figure 36. Relative errors of estimated R and X with regression model $\Delta V \sim I_R + I_X + I_R I_X + I_R^2 + I_X^2 - 1$ without measurement error

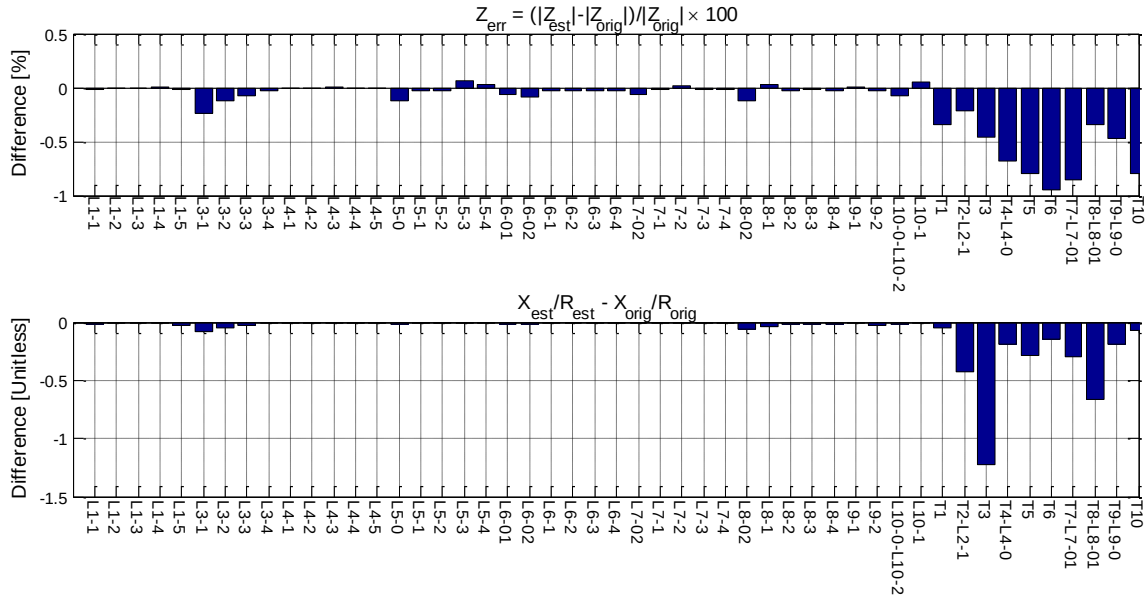

Figure 37. Relative errors of estimated Z and X/R-ratio with regression model $\Delta V \sim I_R + I_X - 1$ without measurement error

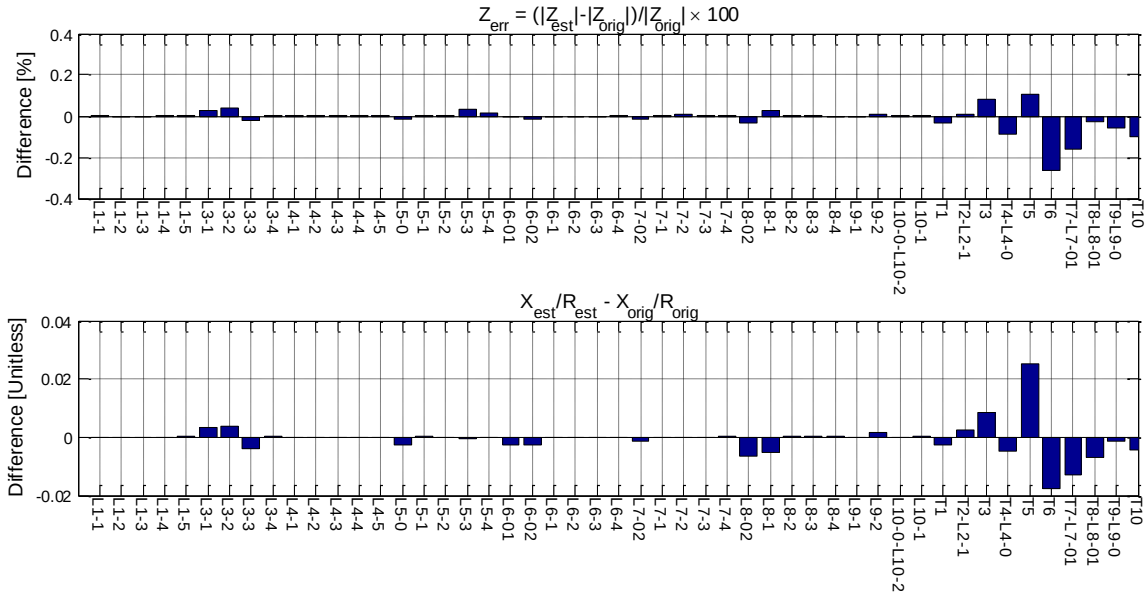

Figure 38. Relative errors of estimated Z and X/R-ratio with regression model $\Delta V \sim I_R + I_X + I_R I_X + I_R^2 + I_X^2 - 1$ without measurement error

4.5.3 Regression Model Comparison with Measurement Error

While the previous section assumes that all voltages and currents are perfectly known, generally any meter has some measurement noise that introduces error. Next, the regression model parameter estimation errors were analyzed with 1% P, 1% Q, and 0.2% V random uniform measurement error. The 0.2% voltage error correspond to the ANSI .2 accuracy class meters [37]. The 1% active and reactive power measurement error level is at or above ANSI .2 and .5 accuracy class meters [37]. This measurement error does not include calibration problems or large bias, and instead is only focused on the stochastic noise common to measurement devices. Table 5 lists the results for all secondary circuit branch R and X parameters estimated with 8759 measurement samples (one year). With measurement error, simpler models perform better than complicated ones. The best overall parameter estimates are obtained with the simplest model $\Delta V \sim I_R + I_X - 1$. Line parameters are estimated best with $\Delta V \sim I_R + I_X - 1$ while transformer resistances are estimated best with model $\Delta V \sim I_R + I_X + I_R^2 - 1$.

Table 5. Relative parameter estimation errors for different linear regression models with 1% P, 1% Q, and 0.2% V measurement error

Included Predictor Variables (All Models Include I_R and I_X)					Avg. Abs. R_{err} [%]			Avg. Abs. X_{err} [%]			Avg. Abs. $\lvert R_{err}\rvert + \lvert X_{err}\rvert$ [%]	Max. R_{err} [%]	Max. X_{err} [%]	Model Order (Best to Worst)
Inter-cept	$I_R \times I_X$	I_R^2	I_X^2	(PF)	Lines	Trafos	All	Lines	Trafos	All				
					2.83	7.38	3.76	3.50	1.51	3.10	6.86	15.30	13.18	1
X					3.08	12.18	4.94	3.48	1.57	3.09	8.03	26.09	13.20	3
		X			5.44	1.18	4.57	3.49	1.60	3.10	7.68	32.61	13.18	2
X		X			14.14	6.42	12.56	3.52	1.60	3.13	15.69	86.10	13.14	4
		X	X		10.40	2.76	8.84	15.80	2.02	12.98	21.82	54.00	43.07	5
X		X	X		16.27	7.95	14.57	15.85	2.01	13.02	27.59	108.13	43.57	8
	X	X	X		10.13	3.85	8.85	20.16	4.01	16.87	25.72	47.77	54.19	6
X	X	X	X		17.19	8.52	15.42	20.23	4.08	16.93	32.35	101.94	54.30	9
		X	X	X	16.25	8.24	14.62	14.80	1.96	12.18	26.79	116.15	60.25	7
X		X	X	X	18.35	9.93	16.63	23.31	5.34	19.64	36.27	118.64	66.16	10
	X	X	X	X	19.68	9.67	17.64	27.54	4.54	22.85	40.48	96.29	85.72	11
X	X	X	X	X	32.84	21.02	30.42	65.05	17.91	55.43	85.86	187.20	225.31	12

The errors of the R and X parameters that are estimated with the regression models $\Delta V \sim I_R + I_X + I_R^2 + I_X^2 + I_R \times I_X - 1$ (best without measurement error) and $\Delta V \sim I_R + I_X - 1$ (best with measurement error) are shown in Figure 39 and Figure 40, respectively. With measurement error, model $\Delta V \sim I_R + I_X + I_R^2 + I_X^2 + I_R \times I_X - 1$ estimates some parameters with considerably higher error than the model $\Delta V \sim I_R + I_X - 1$. This is likely caused by the measurement errors that can be large for the squared and cross-coupling terms. Model $\Delta V \sim I_R + I_X - 1$ estimates most of the line R and X parameters and the transformer X parameters with an acceptable accuracy but does worse in estimating the transformer (branch names that start with a T) resistances.

The errors of the impedance magnitude and X/R-ratio parameters that are estimated with the regression models $\Delta V \sim I_R + I_X - 1$ and $\Delta V \sim I_R + I_X + I_R^2 + I_X^2 + I_R \times I_X - 1$ are shown in Figure 41 and Figure 42, respectively. Excluding the parameters L3-4 and L9-2, $\Delta V \sim I_R + I_X - 1$ estimates all the impedance magnitudes with a good accuracy but

performs worse in estimating the transformer X/R-ratios. $\Delta V \sim I_R + I_X + I_R^2 + I_X^2 + I_R \times I_X - 1$ estimates many impedance magnitudes with considerable error.

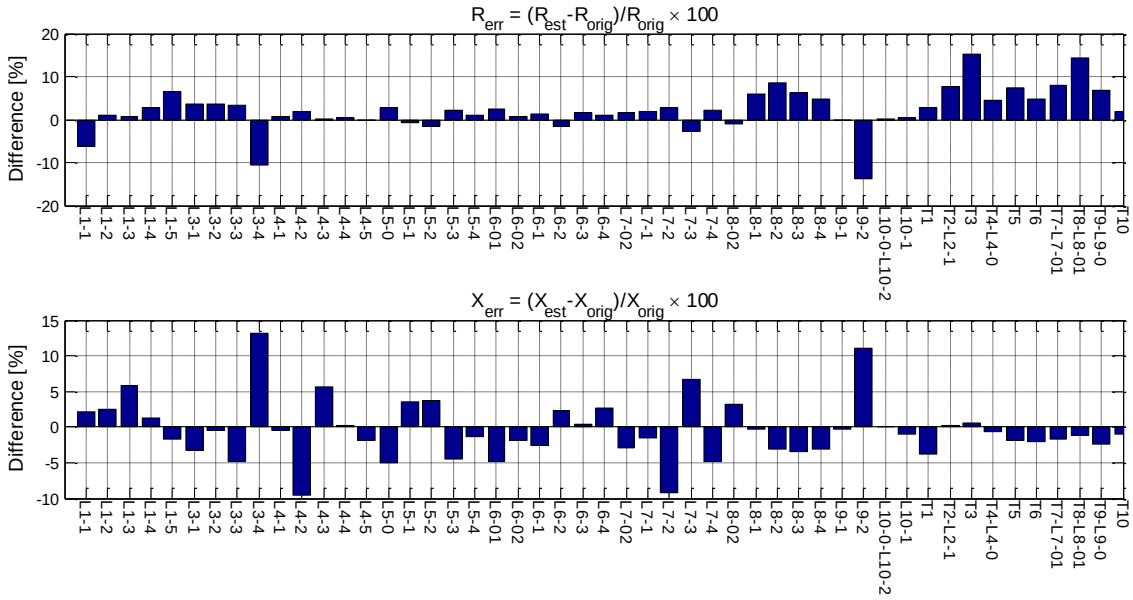

Figure 39. Relative errors of estimated R and X with regression model $\Delta V \sim I_R + I_X - 1$ with 1% P, 1% Q, and 0.2% V measurement error

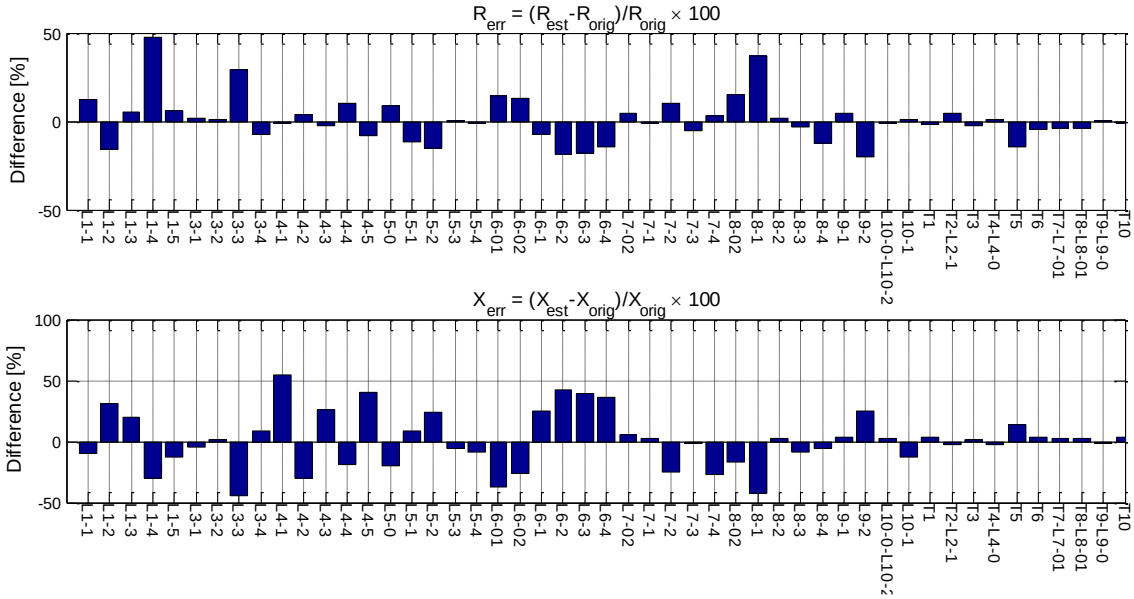

Figure 40. Relative errors of estimated R and X with regression model $\Delta V \sim I_R + I_X + I_R^2 + I_X^2 + I_R \times I_X - 1$ with 1% P, 1% Q, and 0.2% V measurement error

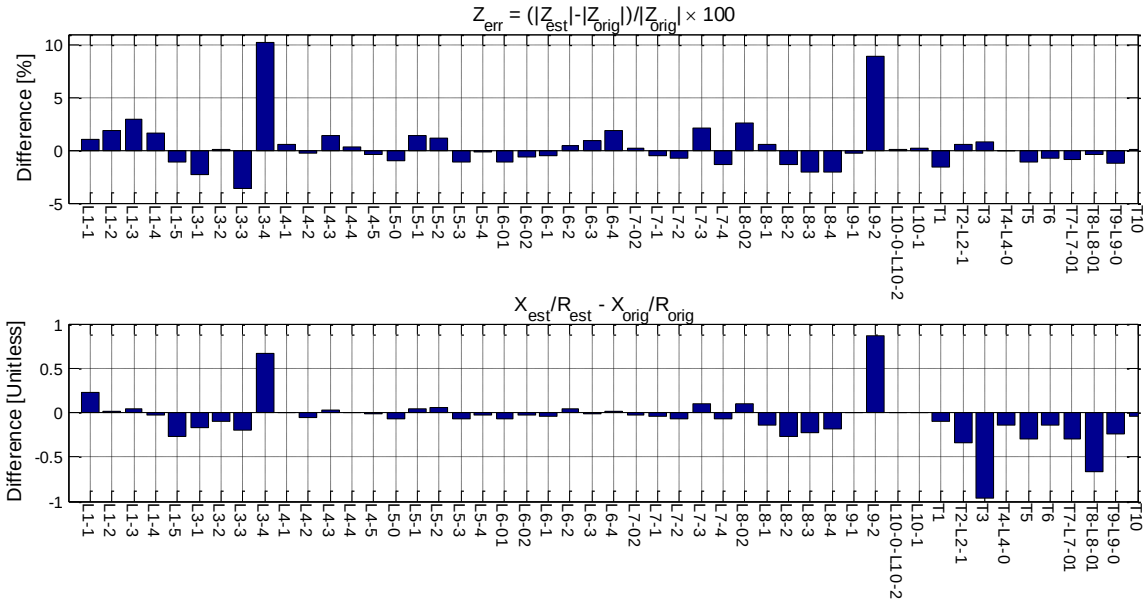

Figure 41. Relative errors of estimated Z and X/R-ratio with regression model $\Delta V \sim I_R + I_X - 1$ with 1% P, 1% Q, and 0.2% V measurement error

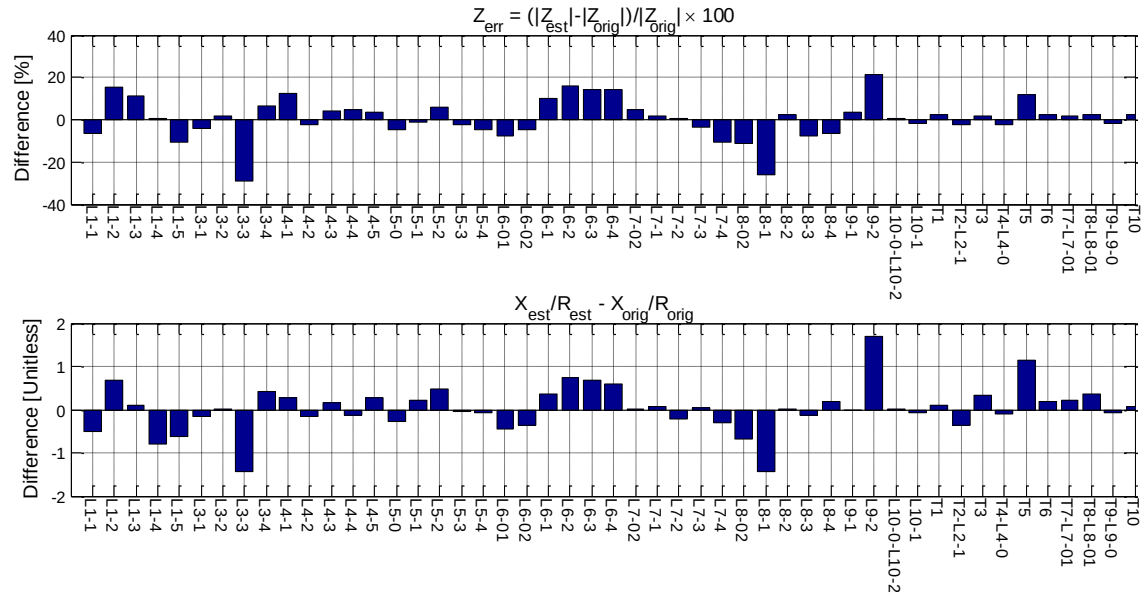

Figure 42. Relative errors of estimated Z and X/R-ratio with regression model $\Delta V \sim I_R + I_X + I_R^2 + I_X^2 + I_R \times I_X - 1$ with 1% P, 1% Q, and 0.2% V measurement error

4.5.4 Adaptive Regression Model Selection

Next, an adaptive regression model selection approach is discussed. The best overall parameter estimation accuracy in the presence of measurement error is obtained

with the simplest model $\Delta V \sim I_R + I_X - 1$. However, other models estimate better service transformer parameters, because the transformer X/R-ratios and impedance magnitudes are higher than those of lines, which results in higher errors in the linearized voltage drop approximation. Based on this insight, this section discusses an adaptive approach where regression problems consisting solely of line parameters are estimated with model $\Delta V \sim I_R + I_X - 1$ and regression problems involving transformer parameters are estimated with a regression model that includes other terms. For this work, the topology is assumed to be known, so each branch is known as either a line or transformer, and all connections between branches are known.

The adaptive approach was analyzed by estimating the transformer R and X parameters in the 66-node test case with 8759 measurement samples. Six different regression models for transformer regression problems were analyzed, and the line parameters were estimated using model $\Delta V \sim I_R + I_X - 1$. The results without measurement error are listed in Table 6, and the results with 1% P, 1% Q and 0.2% V measurement error are listed in Table 7. Similar to section 4.5.2 and 4.5.3, complicated models perform better without measurement error and simple models perform better in the presence of measurement error. The best overall parameter estimates were obtained with model $\Delta V \sim I_R + I_X + I_R^2 - 1$, i.e., a model that includes the squared current term I_R^2.

Figure 43 and Figure 44 compare the performance of the adaptive approach with (transformer regression problem) models $\Delta V \sim I_R + I_X - 1$, $\Delta V \sim I_R + I_X + I_R^2 - 1$, and $\Delta V \sim I_R + I_X + I_R^2 + I_X^2 - 1$ with different P, Q, and V error levels and sample sizes. Model $\Delta V \sim I_R + I_X - 1$ is equal to the nonadaptive approach with regression model $\Delta V \sim I_R + I_X - 1$. Model $\Delta V \sim I_R + I_X + I_R^2 - 1$ beats model $\Delta V \sim I_R + I_X - 1$ as long as sufficiently large sample size is used. As shown in Figure 44, model $\Delta V \sim I_R + I_X + I_R^2 - 1$ outperforms model $\Delta V \sim I_R + I_X + I_R^2 + I_X^2 - 1$ independent of the error level and sample size.

Table 6. Relative parameter estimation errors of the adaptive approach without measurement error

Included Predictor Variables for Regression Problems with Transformer Parameters (All Models Include I_R and I_X)					Avg. Abs. R_{err} [%]			Avg. Abs. X_{err} [%]			Avg. Abs. $\lvert R_{err}\rvert + \lvert X_{err}\rvert$ [%]	Max. R_{err} [%]	Max. X_{err} [%]	Model Order (Best to Worst)
Inter-cept	$I_R \times I_X$	I_R^2	I_X^2	(PF)	Lines	Trafos	All	Lines	Trafos	All				
					0.517	7.833	2.010	0.349	1.471	0.578	2.588	18.548	2.671	3
		X			0.517	1.384	0.694	0.349	1.593	0.603	1.297	2.638	2.782	4
			X		0.517	12.371	2.937	0.349	9.506	2.218	5.154	31.464	14.792	6
		X	X		0.517	0.439	0.501	0.349	0.575	0.395	0.896	2.638	1.218	2
	X				0.517	7.903	2.025	0.349	6.657	1.636	3.661	18.615	9.563	5
	X	X	X		0.517	0.186	0.450	0.349	0.100	0.298	0.748	2.638	1.218	1

Table 7. Relative parameter estimation errors of the adaptive approach with 1% P, 1% Q, and 0.2% V measurement error

Included Predictor Variables for Regression Problems with Transformer Parameters (All Models Include I_R and I_X)					Avg. Abs. R_{err} [%]			Avg. Abs. X_{err} [%]			Avg. Abs. $\lvert R_{err}\rvert + \lvert X_{err}\rvert$ [%]	Max. R_{err} [%]	Max. X_{err} [%]	Model Order (Best to Worst)
Inter-cept	$I_R \times I_X$	I_R^2	I_X^2	(PF)	Lines	Trafos	All	Lines	Trafos	All				
					2.834	7.380	3.761	3.500	1.514	3.095	6.857	15.299	13.178	4
		X			2.834	1.009	2.461	3.500	1.603	3.113	5.574	13.667	13.178	1
			X		2.834	11.809	4.665	3.500	9.293	4.683	9.348	26.657	15.161	6
		X	X		2.834	2.878	2.843	3.500	2.108	3.216	6.059	13.667	13.178	2
	X				2.834	7.444	3.774	3.500	6.547	4.122	7.897	15.339	13.178	5
	X	X	X		2.834	3.274	2.924	3.500	3.205	3.440	6.364	15.846	13.178	3

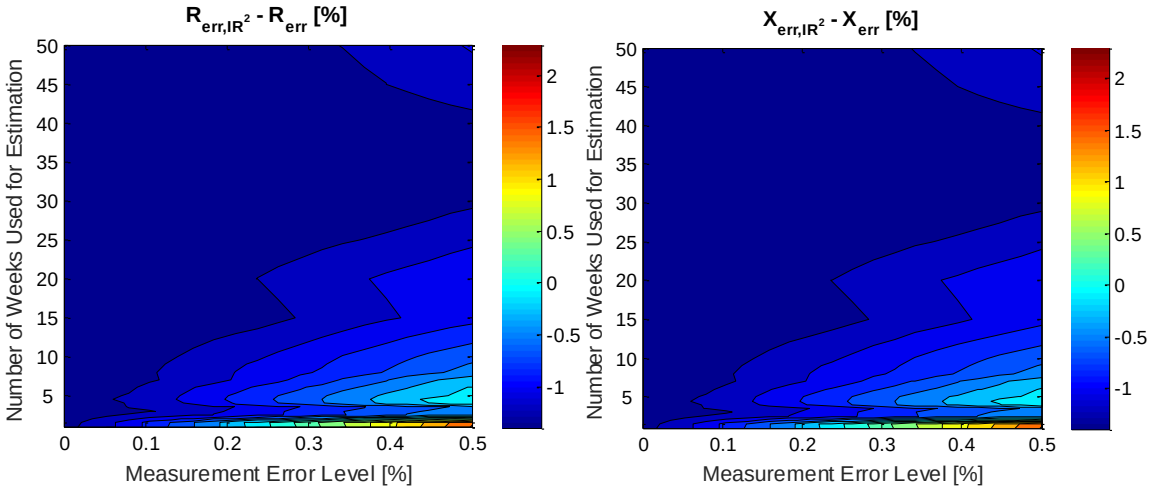

Figure 43. Difference of relative errors of estimated R and X between adaptive regression model $\Delta V \sim I_R + I_X + I_R^2 - 1$ and the non-adaptive model $\Delta V \sim I_R + I_X - 1$

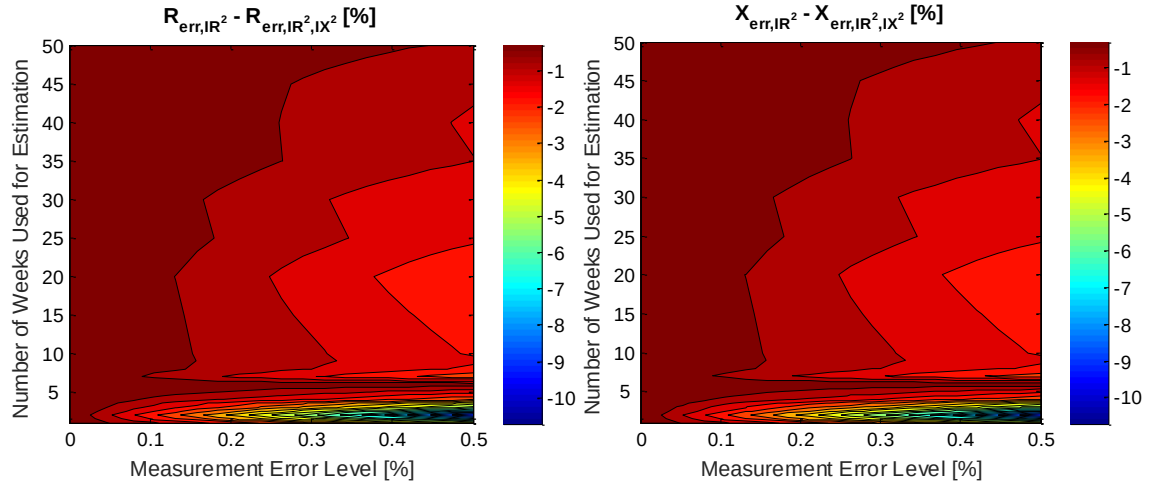

Figure 44. Difference of relative errors of estimated R and X between regression models $\Delta V \sim I_R + I_X + I_R^2 - 1$ and $\Delta V \sim I_R + I_X + I_R^2 + I_X^2 - 1$

Based on these results shown above, the rest of this chapter utilizes the adaptive parameter estimation approach where regression models $\Delta V \sim I_R + I_X - 1$ and $\Delta V \sim I_R + I_X + I_R^2 - 1$ are used for regression problems without and with transformer parameters, respectively. Thus, the regression models presented in section 4.2 are used for circuit subsections without transformers. On the other hand, single transformer parameters are best estimated by utilizing design matrix $\boldsymbol{\mathcal{X}}$ and parameter vector $\boldsymbol{\beta}$ given by

$$\boldsymbol{\mathcal{X}} = [\boldsymbol{I}_{R1}, \boldsymbol{I}_{X1}, \boldsymbol{I}_{R1}^2] \tag{37}$$

$$\boldsymbol{\beta} = \left[R_1, X_1, \beta_{Rsq,1}\right]^{\mathrm{T}} \tag{38}$$

where parameter $\beta_{Rsq,1}$ does not have a direct physical meaning. The response vector $\boldsymbol{y}$ is the same as in (23). Similarly, the parameters of N parallel branches, i^{th} of which is a transformer, are best estimated by utilizing design matrix $\boldsymbol{\mathcal{X}}$ and parameter vector $\boldsymbol{\beta}$ are given by

$$\boldsymbol{\mathcal{X}} = \begin{bmatrix} \mathbf{I} & [-\boldsymbol{I}_{R,1} & -\boldsymbol{I}_{X,1}] & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & & \ddots & \vdots & \vdots \\ \mathbf{I} & \mathbf{0} & & \cdots & [-\boldsymbol{I}_{R,N} & -\boldsymbol{I}_{X,N}] & -\boldsymbol{I}_{R,i}^2 \end{bmatrix} \tag{39}$$

$$\boldsymbol{\beta} = \left[V_{0,1}, \dots, V_{0,M}, R_1, X_1, \dots, R_N, X_N, \beta_{Rsq,i}\right] \tag{40}$$

The response vector $\boldsymbol{y}$ is the same as in (32).

These selected regression models are optimized for the practical setting where the measurement error dictates the parameter estimation accuracy. Without measurement error, parameters can be estimated with a smaller error by using regression models with additional higher-order terms of the predictor variables.

4.6 Three-Phase Test Circuit Parameter Estimation Results

This section presents detailed parameter estimation results for the three-phase test circuit for different measurement sample sizes and different levels of power and voltage measurement error.

4.6.1 Results with Respect to Sample Size

Typically, the utilities have abundant AMI measurement data that can be filtered for the offline parameter estimation. However, in order to keep the parameter estimation algorithm computationally and data efficient, the necessary number of measurement samples must be determined. A sufficiently large measurement sample should be selected to achieve the highest possible accuracy, but at some point adding more samples is

expected to have diminishing marginal returns. The parameter estimation error dependency on the number of measurement samples was studied on the 66-node test circuit. Figure 45 shows the relative errors of the estimated R and X parameters with measurement sample lengths 1-53 weeks when no measurement error is present. The measurement samples were selected in a random order from the available set of 53 weeks of load data. Each plot contains 49 lines, one for each R or X parameter. The average error of all parameters is shown with the red bold line. As the figure indicates, when no measurement error is present, there is no need to utilize large numbers of measurement samples. It should also be noted that utilizing a larger number of measurement samples can also reduce parameter estimation accuracy depending on the characteristics of the additional samples. Generally, parameter estimation results depend on the characteristics of the utilized measurement data and measurement error.

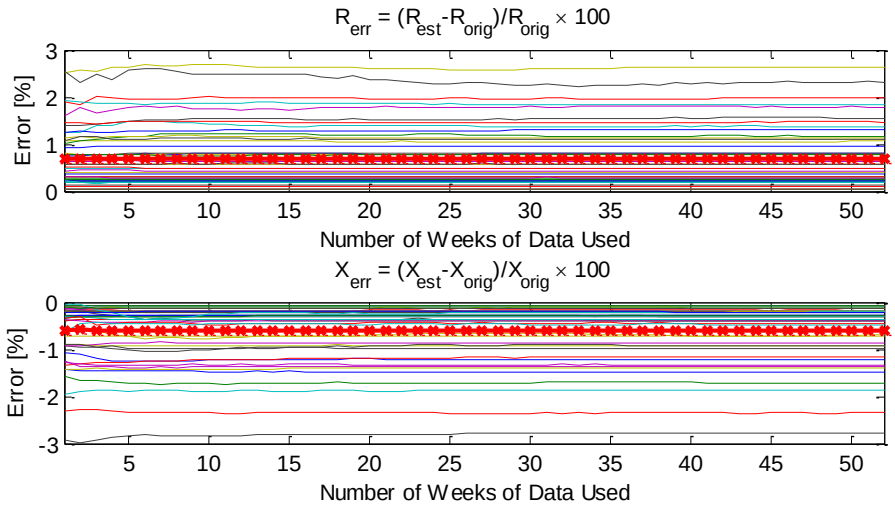

Figure 45. Relative errors of the estimated R and X with different measurement sample sizes without measurement error

The detailed parameter estimation accuracies without measurement error are shown in Figure 46. All the parameters are estimated with an error less than 3%. As discussed in section 4.5, considerably better parameter estimation accuracies without measurement error are achieved with alternative regression models. However, the proposed adaptive

regression model approach achieves the best performance in the practical setting with noisy measurements.

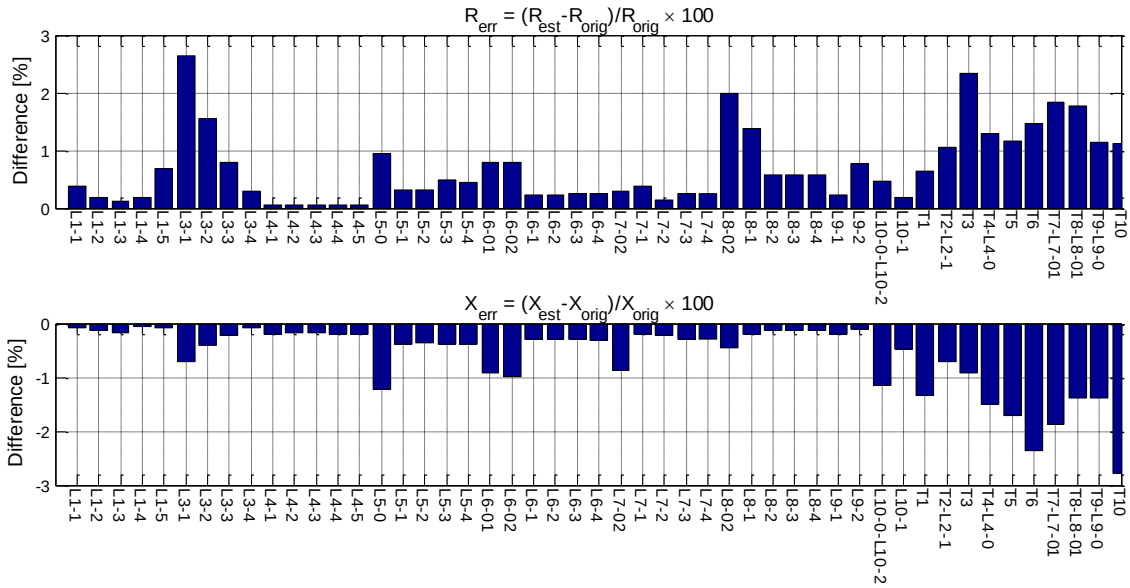

Figure 46. Relative errors of the estimated R and X with 53 weeks of measurement data and without measurement error

4.6.2 Results with Power Measurement Error

Next, the parameter estimation error dependency on P and Q measurement error and sample size was studied. Measurement error was added to each active and reactive power measurement sample i with $P_i = (1+\delta)P_i$ and $Q_i = (1+\delta)Q_i$, where the measurement error magnitude was set to $\delta \in \text{Uniform}(-0.05, 0.05)$. Perfect voltage measurements were assumed.

Figure 47 shows the average absolute R and X estimation errors of all the 66-node test circuit parameters with measurement sample lengths from 1 through 50 weeks with different P and Q measurement error levels and sample sizes. With reasonably small measurement error levels and sufficient sample sizes, the average parameter estimation errors are well below 1-2%. In the presence of measurement error, increasing the measurement sample size considerably improves the accuracy of the estimated parameters

up to the sample sizes of around 10 weeks (1680 samples) after which adding further samples has only small if any improvement. The estimation accuracy does not improve monotonically with the sample size due to the randomness of the load data and the sample selection. If Figure 47 was repeatedly plotted over a randomly drawn order of the load data, the average error of the repetitions is expected to reduce monotonically as the sample size grows.

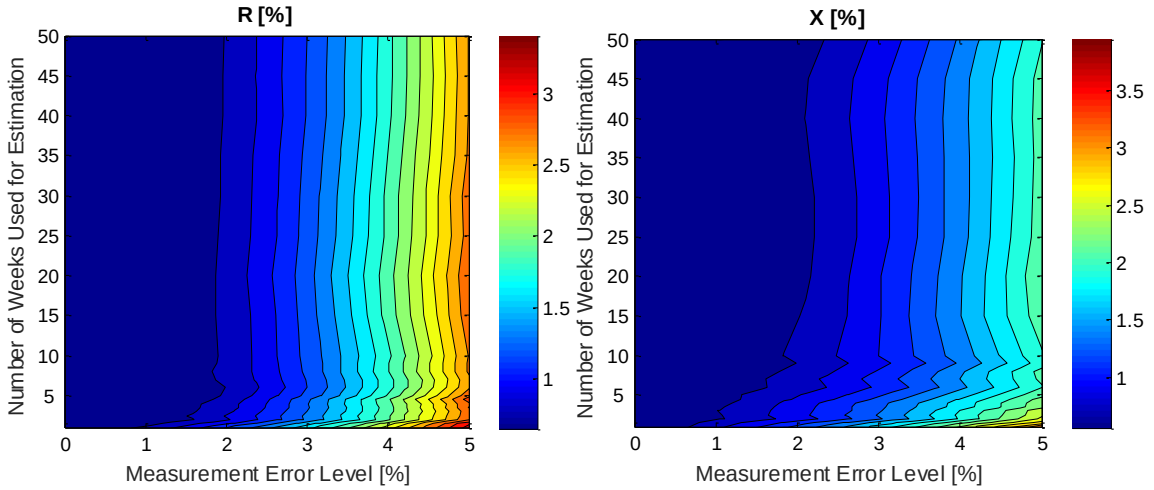

Figure 47. Average relative errors of R (left) and X (right) estimated with 1-50 weeks of load data and 0-5% of P and Q measurement error

4.6.3 Results with Voltage Measurement Error

Next, parameter estimation performance was studied in the presence of voltage measurement error. The same principle was used as with the power measurement error above, but now voltage measurement error up to 0.5% (e.g. Class 0.5 smart meter) was added to the voltage measurements. Figure 48 shows the average absolute R and X estimation errors of all the 66-node test circuit parameters with measurement sample lengths from 1 through 50 weeks with different voltage measurement error levels and sample sizes. Again the errors of the estimated parameters reduce (although not monotonically) as the sample size is grown. Clearly, voltage measurement error has a much

larger influence on the parameter estimation accuracy than the power measurement error. Therefore, it is an imperative to have high quality voltage measurements.

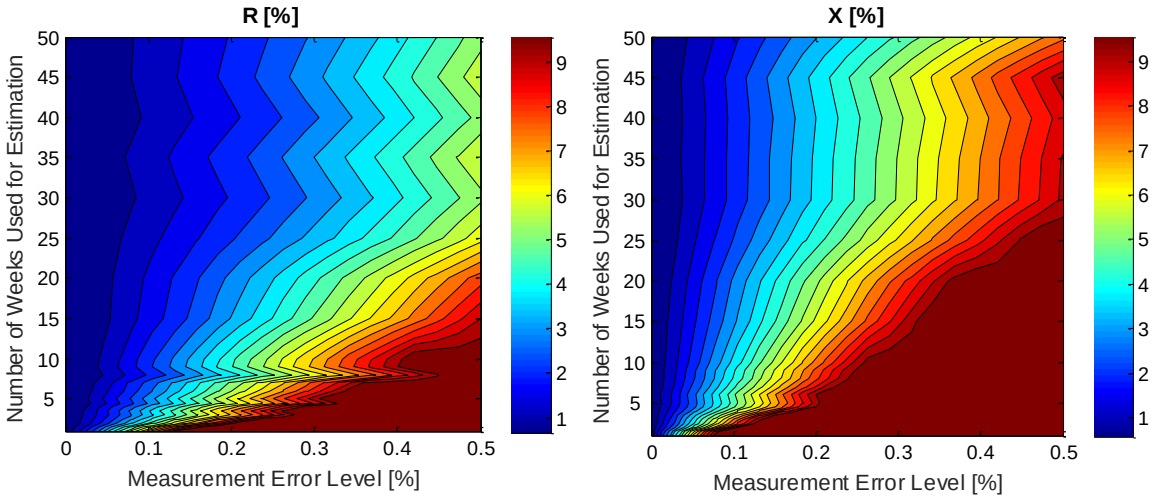

Figure 48. Average relative errors of R (left) and X (right) estimated with 1-50 weeks of load data and 0-0.5% of V measurement error, error magnitudes >10% are set to 10%

It is relevant to point out the reason for the parameter estimation to be very sensitive for the voltage measurement error. Figure 49 shows the boxplots of load voltage measurement errors at error level 0.2%, voltage drops over the service lines, and voltage drops over the service transformer for the entire year of load data (8760 samples). Each boxplot on the top of Figure 49, represents the load voltage measurement errors. Each boxplot in the bottom of Figure 49 represents the voltage drops over a given branch or transformer. The voltage drops over some lines are on the same order or smaller than the measurement error. Since the proposed parameter estimation utilizes the branch voltage drop as the linear regression response variable, it is not possible to estimate effectively impedance parameters for branches over which the voltage drop is less or equal than the voltage measurement error level. Therefore, it is imperative to have high quality voltage measurements.

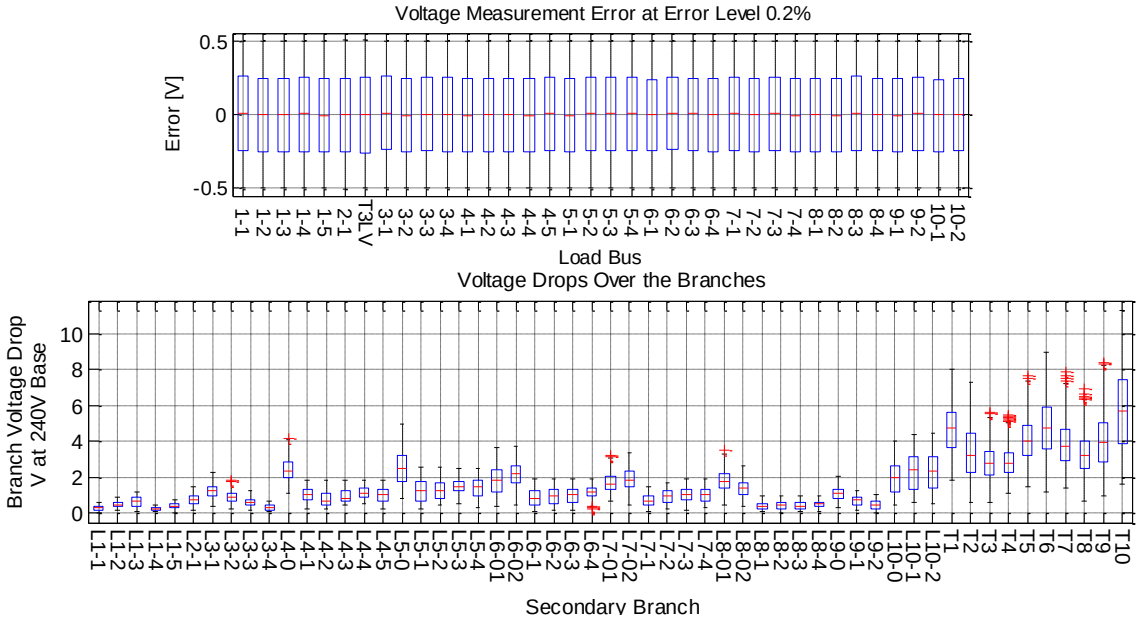

Figure 49. Load voltage measurement errors at error level 0.2% (top) compared to voltage drops over the 240V base secondary circuit transformers and lines (bottom)

4.6.4 Results with Power and Voltage Measurement Error

Finally, parameter estimation performance is shown in the presence of both power and voltage measurement error. Similarly to the previous subsections, measurement error up to 0.5% (e.g. Class 0.5 smart meter) was added to the P, Q, and V measurements. Figure 50 shows the average absolute R and X estimation errors of all the 66-node test circuit parameters with measurement sample lengths from 1 through 50 weeks with different voltage measurement error levels and sample sizes. Again the errors of the estimated parameters decrease (although not monotonically) as the sample size is grown. Even at relatively high measurement error levels of 0.5%, the average absolute error or the estimated R and X can be brought down to around 6% and 9%, respectively by utilizing sufficiently large sample sizes. However, adding more samples does not completely remove the influence of measurement error. Therefore, it is necessary to have high-quality (especially voltage) measurements in order to accurately estimate the parameters.

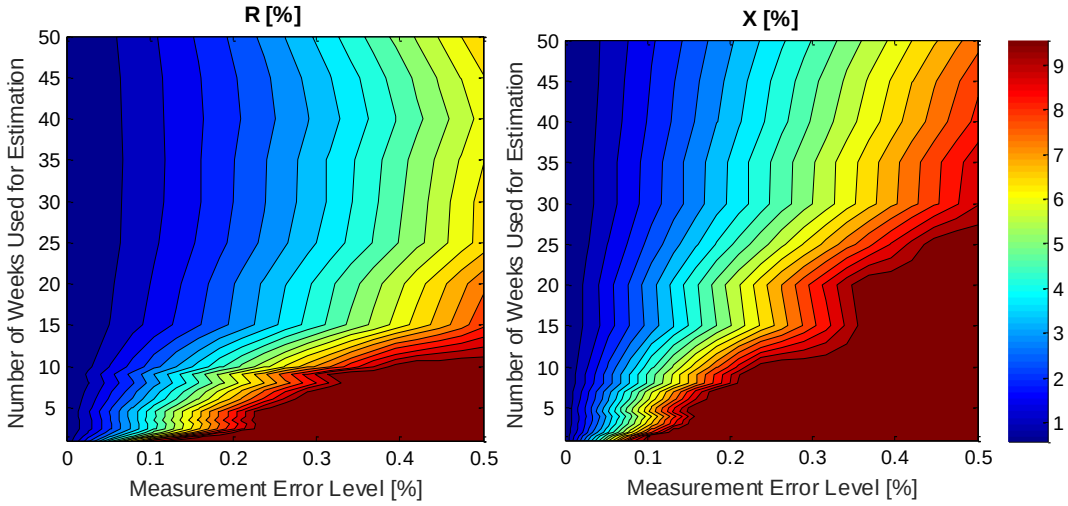

Figure 50. Average absolute R (left) and X (right) estimation errors for 1-50 weeks of load data and 0-0.5% of P, Q, and V measurement error, error magnitudes >10% are set to 10%

The relative errors of the estimated R and X parameters with 0.2% measurement error level and 8759 measurement samples are shown in Figure 51. Excluding parameters of L3-4 and L9-2, all the parameters are estimated with a reasonably good accuracy with mean (maximum) relative error of R and X at 2.05% (8.67%) and 2.73% (9.50%), respectively. The relative errors of the estimated Z and absolute errors of the estimated X/R-ratios are shown in Figure 52. Again, excluding parameters of L3-4 and L9-2, the mean (maximum) relative Z and absolute X/R-ratio errors were 1.10% (3.61%) and 0.08 (0.28), respectively.

No straightforward reason was found for the low quality of L3-4 and L9-2 parameter estimates. The voltage drop over branch L3-4 is quite small but not uniquely small in the test circuit. Branch L9-2 has a relatively high X/R-ratio compared to the other branches in the test circuit. However, neither of these factors fully explains the considerably poorer quality of the estimated parameters. Additionally as shown in Figure 53 the two regression problems that contain L3-4 and L9-2 have very low R-squared values, which indicates a low quality fit. However, there are other well-estimated parameters with similar R-squared values and thus, R-squared cannot be directly used to

describe the quality of a regression model. Figure 53 also shows that the sum of squared errors of these two regression problems are not considerably higher and that the parameter p-values are significant.

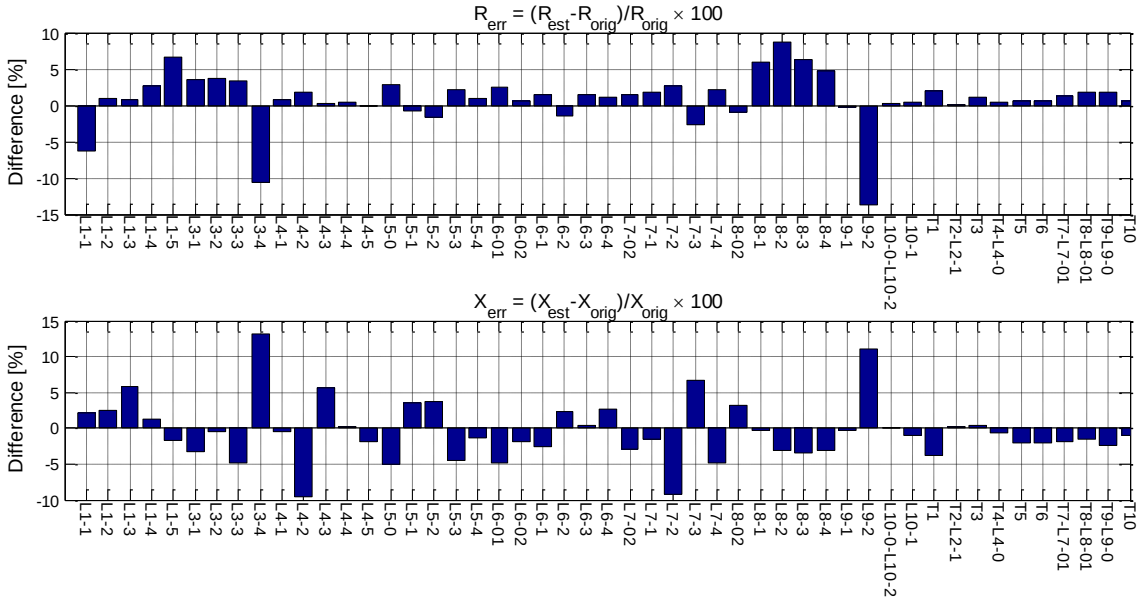

Figure 51. Relative errors of estimated R and X with 1% P, 1% Q, and 0.2% V measurement error when the parameters are estimated with the adaptive approach

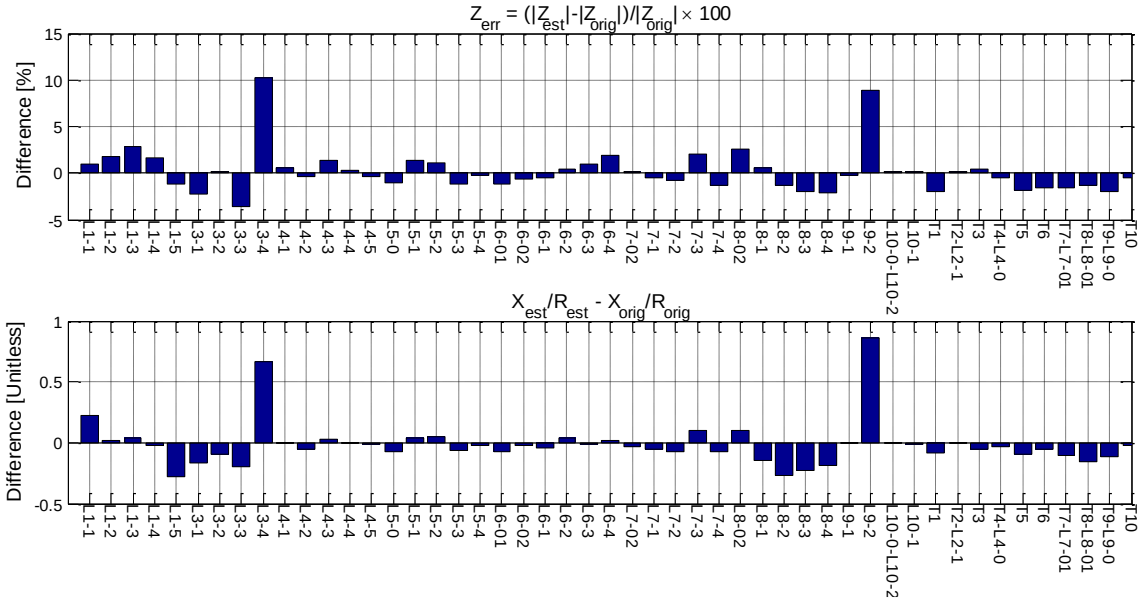

Figure 52. Relative errors of estimated Z and X/R-ratio with 1% P, 1% Q, and 0.2% V measurement error when the parameters are estimated with the adaptive approach

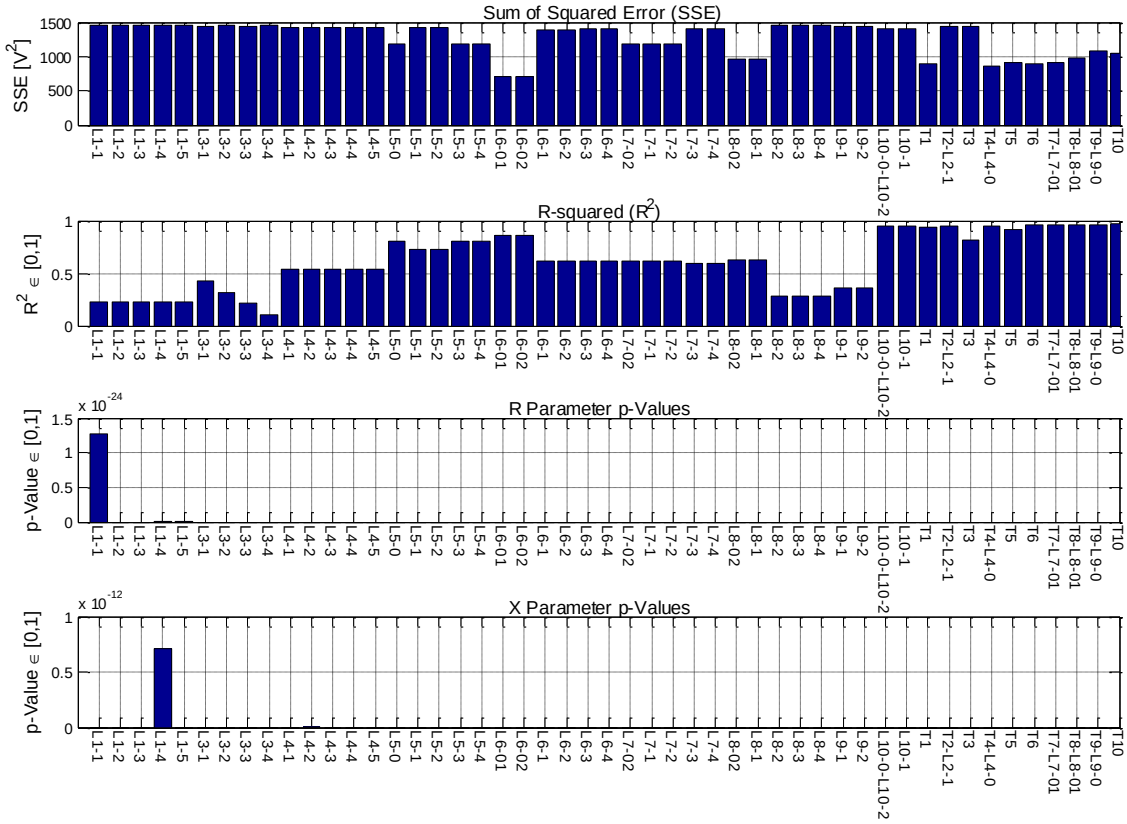

Figure 53. Sum of squared errors, R-squared values, R and X p-values with 1% P, 1% Q, and 0.2% V measurement error

Finally, as Figure 54 and Figure 55 indicate, the regression problems that include L3-4 and L9-2 do not have particularly high or low means and/or standard deviations of the response and/or predictor variables. Potentially, other characteristics in the load data of L3-4 and L9-2 would explain the lower estimation quality of these parameters. To conclude, more work is needed to determine the approach to detect and fix parameter estimation regression problems with low quality.

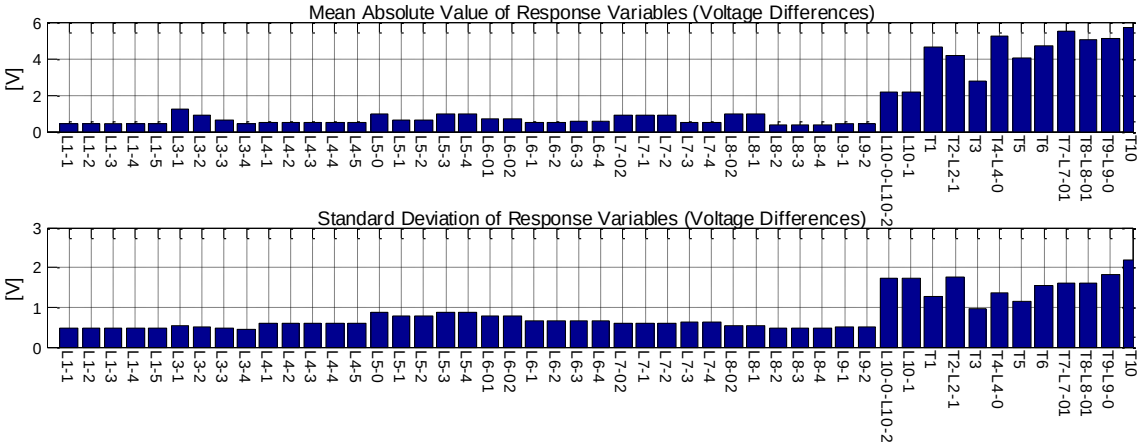

Figure 54. Means and standard deviations of the response variables (voltage drops)

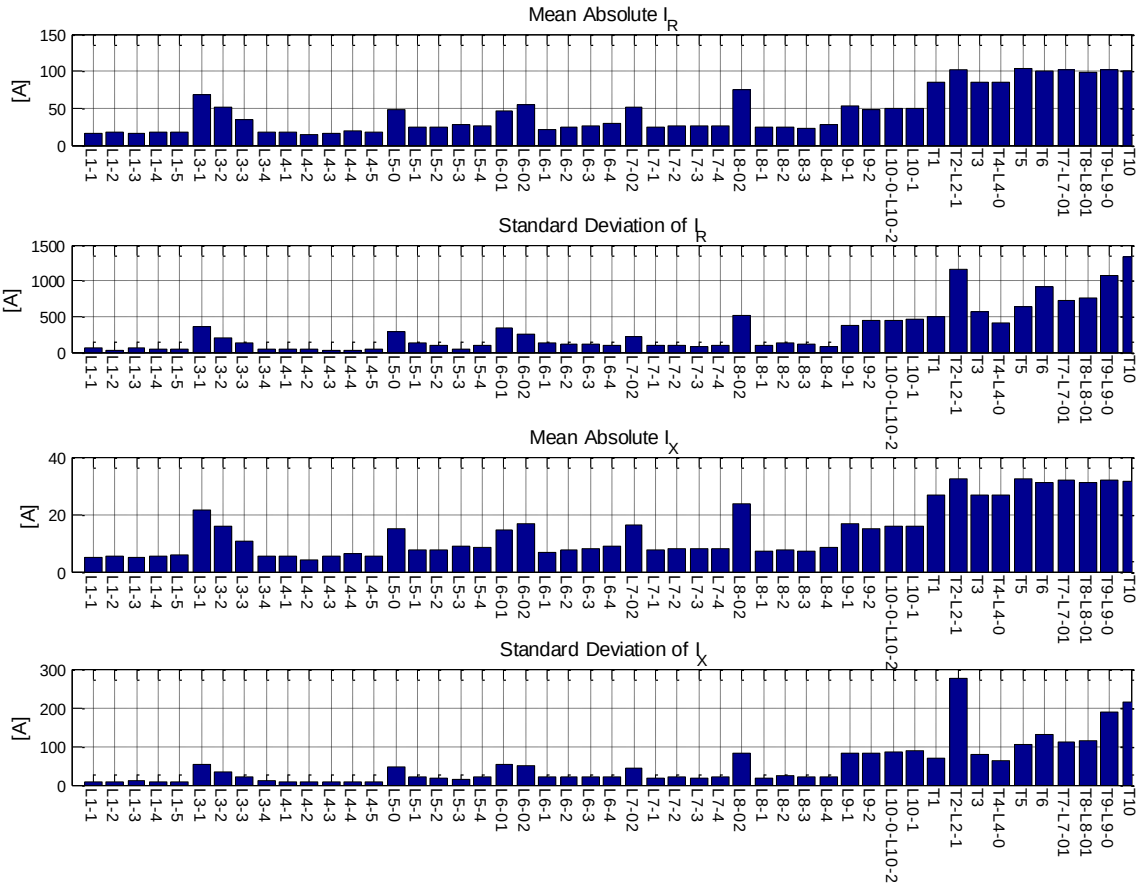

Figure 55. Means and standard deviations of the predictor variables I_R and I_X

Figure 56 shows the absolute and relative errors of the simulated per-unit voltage drops from the transformer primary winding to the load buses. The errors are calculated

between the voltages simulated with the true parameters and the voltages simulated with the estimated parameters. In both cases, the voltages were simulated with the true P and Q values. All the errors are so small that in real circuits, they can be hard to distinguish from measurement noise and other modeling inconsistencies.

It should be noted that since parameter estimation results are data-driven, different results are obtained with different characteristics of load data and measurement error. As shown in 4.5.2, without measurement error parameters can be estimated with a very small error especially if regression models with additional terms are used. The presented parameter estimation approach is optimized for the practical setting where the measurement error dictates the parameter estimation accuracy.

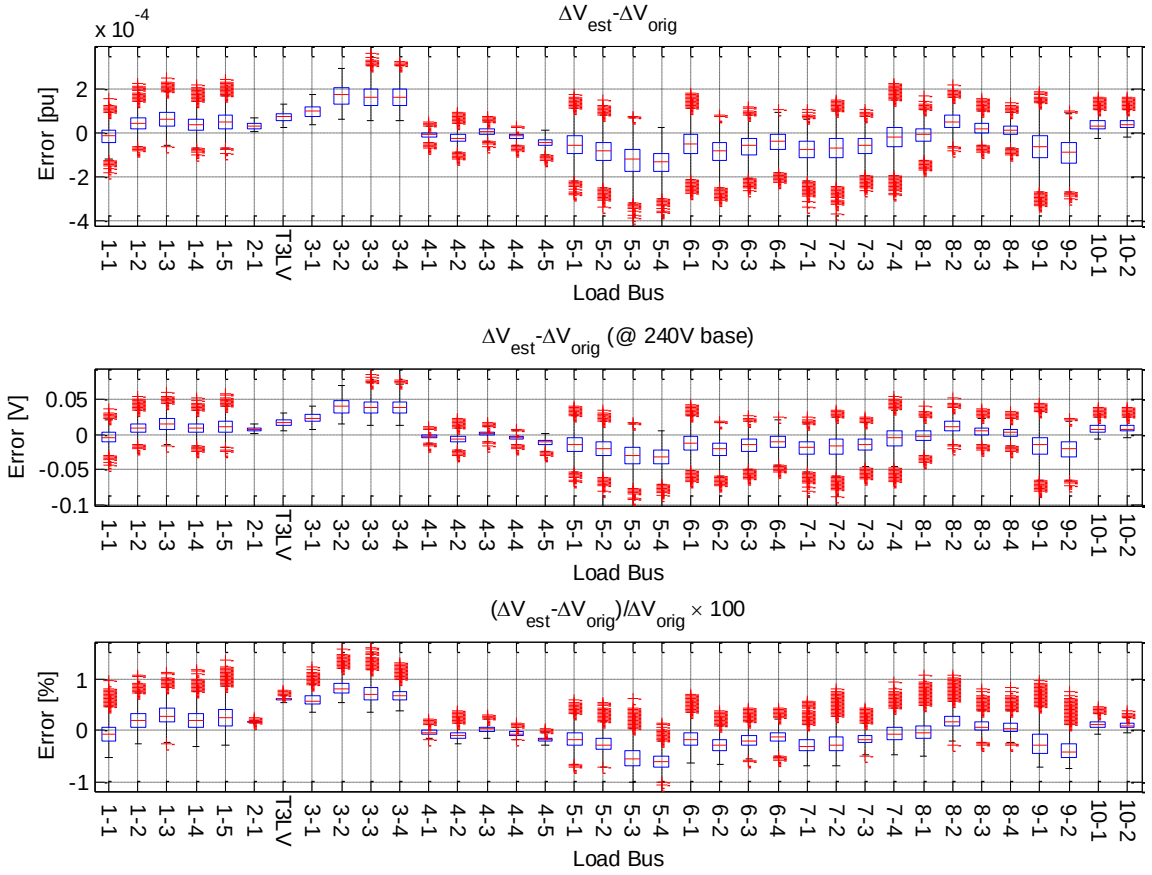

Figure 56. Errors of simulated voltage drops from the service transformer primary to the load buses when the parameters are estimated with the adaptive approach

4.7 Georgia Tech Feeder Results

The proposed parameter estimation algorithm was utilized to calibrate the secondary circuit parameters of one of the Georgia Tech feeders. Since the true parameter values are unknown, the parameter estimation accuracy was measured with mean bias errors (MBE $= \frac{1}{n}\sum_{i=1}^{n}\frac{(V_{sim}-V_{meas})}{V_{meas}}$) of the voltage drops simulated with the basecase parameters and the estimated parameters. The results are shown in Figure 57. The parameter estimation very effectively reduced the bias of the voltage drop simulation error.

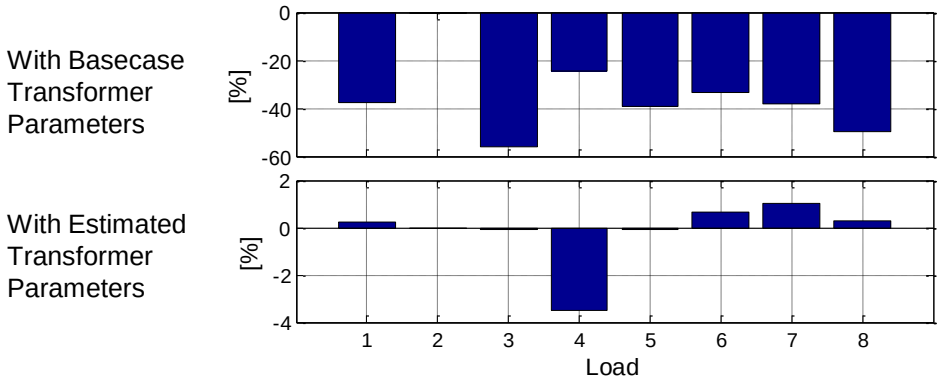

Figure 57. Mean bias errors of simulated voltage drops from the substation with base case transformer parameters (top) and estimated transformer parameters (bottom)

Typically distribution system secondary circuits are fed by a single service transformer whose upstream and downstream bus voltages are rarely measured. In order to estimate the transformer impedance, voltage estimates of both of the transformer buses are needed. The downstream bus voltages can be estimated from the secondary circuit measurements with the hierarchical radial circuit parameter estimation approach shown above. However, the same approach cannot be used to estimate the transformer upstream bus voltages. Instead, the transformer upstream bus voltages can be estimated from time series power flow analysis by assuming that the distribution system primary circuit is well-modeled and any secondary system impedance inaccuracies have only a small impact on the transformer medium-voltage side voltages.

It is challenging to verify the Georgia Tech primary system model accuracy due to the absence of measurements between the loads in the secondary circuits and the substation. There is uncertainty with respect to the primary system underground cable capacitances and in some cases the exact primary system topology. Moreover, the Georgia Tech AMI historical database has errors, many of which have already been detected and removed but further undetected problems are likely to exist [42], [91].

Figure 58 shows the boxplots of the relative voltage simulation errors for all the 10 secondary circuit loads in the studied Georgia Tech feeder. The larger mean and range of errors are explained next. Meter 136E_ML1 records instantaneous measurements, which are subject to considerable variation in a given 15-min measurement time period. Since the other feeder measurements are 15-min averages, these instantaneous measurements do not synchronize well, so simulated voltages have more variation. Due to large research equipment in the building, meter B149E_MH2 has an abnormal load shape that is only at a few kW most of the time and occasionally jumps quickly up to 200-300kW for a while. Filtering out these jumps from the parameter estimation improves the estimated parameters and reduces the errors for test data periods without the jumps. However, voltage simulation errors during these jumps cannot be eliminated. Building B199E_MH1 has a lot of PV with negative power injections present during the daytimes. Parameter estimation effectively reduced the means and variations of the other loads.

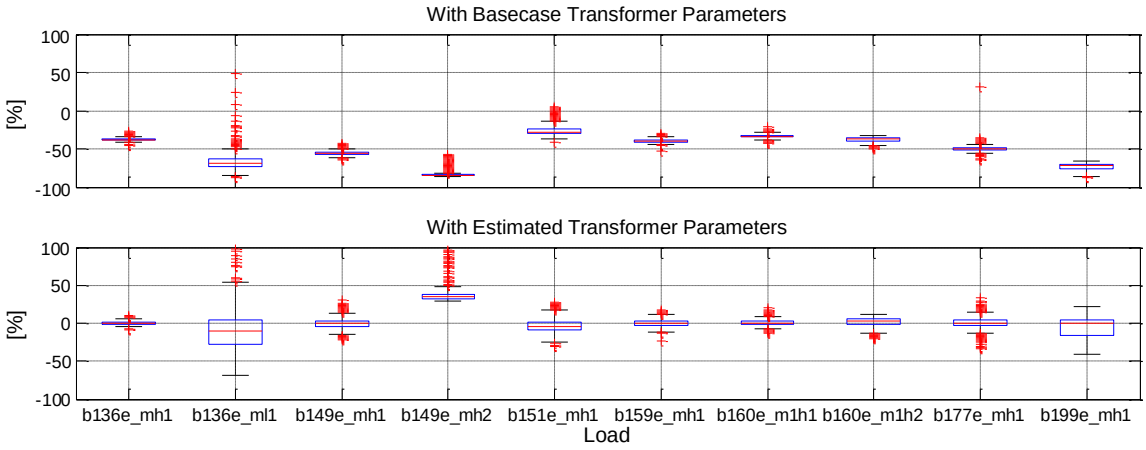

Figure 58. Relative voltage drop simulation errors with the basecase transformer parameters (top) and estimated transformer parameters (bottom)

4.8 Discussion

The presented LRPE method assumes a well-modeled primary circuit, known secondary circuit topologies, and AMI active power, reactive power, and voltage measurements at all the loads in the secondary circuit. No existing secondary circuit model information is needed, except for topology. This makes the proposed method very applicable to the current industry needs. The next chapter presents parameter estimation methods that can be used when there are meters not reporting voltages or when the available measurements are much more limited.

It is worth noting that the proposed linear regression parameter estimation algorithm requires using the hierarchical principle shown in Algorithm 2. It is possible to formulate a large linear regression problem for estimating all the branch impedances at once but unfortunately, the resulting design matrix $\boldsymbol{X}$ is perfectly collinear and has no unique solution. The reason for this is that all upstream branch predictors are linear combinations of the downstream predictors. The regression problem can still be solved, but it is unclear how to set sufficient additional conditions to get a unique solution that would provide the branch parameter estimates. Alternatively, the predictor linear dependency could be avoided by utilizing a nonlinear relationship between the voltage drop

and the downstream predictors (such as the AC power flow). However, the resulting problem would not be linear with respect to the parameters, and iterative nonlinear optimization algorithms would be needed.

The best linear regression model depends, among other things, on the characteristics of the data and on the true values of the impedance parameters. More complicated models can better estimate true impedance values under conditions of low measurement noise levels. On the other hand, the higher the measurement error level is or the lower the true impedance magnitudes (and the voltage drop) are, the simpler regression models should be used. The proposed adaptive linear regression model selection method is a necessary compromise between accuracy and insensitivity to measurement error.

The performance of the LRPE method is demonstrated on a three-phase test circuit with ten different secondary circuit topologies and on the Georgia Tech campus distribution system with AMI data. The parameter estimation accuracy depends on the sample size and error level of the measurement data utilized for the parameter estimation. Large sample sizes, which LRPE method can easily handle, should be used to minimize the impact of measurement error. At practical measurement error levels, the three-phase test circuit secondary circuit parameters can be estimated with an average error less than 3%. However, high-quality voltage measurements are needed to accurately estimate small impedance values. In the Georgia Tech feeder, the parameters estimated with LRPE resulted in considerably lower voltage simulation errors. To accurately estimate utility feeder parameters, gross errors must be removed from the primary system model and the measurement data.

CHAPTER 5. DISTRIBUTION SECONDARY CIRCUIT PARAMETER ESTIMATION WITH LIMITED MEASUREMENTS

The LRPE parameter estimation method presented in Chapter 4 required that AMI active power, reactive power, and voltage measurements be available at all loads. This chapter extends the distribution system parameter estimation scope to cases where the available measurement data is more limited. First, section 5.1 extends the LRPE algorithm from Chapter 4 to handling meters that do not report voltage measurements. Then, section 5.2 proposes a parameter estimation method in the currently common case when a utility does not have a dense network of smart meters (or other sensors) in the secondary circuits. Most of the work presented in this chapter has been published in [97]–[99], [107].

5.1 Estimation with Meters Not Reporting Voltages

Smart meters and PV micro-inverters measure both voltages and currents to derive power measurements from them. Modern smart meters can allow firmware upgrades [108] to be able to transmit the voltage measurements to the utility database. However, older smart meters may not have either of these capabilities and thus, in practice some meters may provide only power (or current) measurements. This section presents a modified secondary circuit parameter estimation method that can handle a limited amount of meters that do not transmit voltage measurements. It should be noted that any meter without voltage measurements reduces the accuracy and observability of the parameter estimation and thus, it is desirable to have high-quality voltage measurements from all smart meters.

The modified method presented in this section relies on the assumptions listed in section 4.1 except that all leaf nodes without voltage measurements are assumed to have current and power factor measurements. If a meter without voltage measurements has only power measurements, these power measurements must be converted to current

measurements (I_R and I_X) by utilizing estimated (e.g. nominal or simulated) voltages. Once this has been done, the modified algorithm proceeds as described in Algorithm 3 to estimate the parameters of each secondary circuit.

Algorithm 3: LRPE algorithm with meters not reporting voltages

Input: A list of all secondary circuit branches $\mathcal{L}$ with fields upstream and downstream node, number of parallel branches, branch current measurements $\boldsymbol{I}_R, \boldsymbol{I}_X$, branch node voltage measurements $\boldsymbol{V}$
Output: Secondary circuit branch estimation results including the estimated parameters R_{est}, X_{est} and their p-values R_{pval}, X_{pval}, and the regression model R^2- and MSE-values
Initialization: Set the list of active branches, ℓ, empty and remove from $\mathcal{L}$ the upstream branches of leaf nodes without voltage measurements (their parameters cannot be estimated). Add the currents of such leaf nodes to the currents of their parent nodes.
1. IF $\mathcal{L}$ is empty, **STOP**.
ELSEIF $\mathcal{L}$ has a branch i with both upstream and downstream voltage measurements or estimates, set $\ell = i$, remove ℓ from $\mathcal{L}$.
ELSEIF $\mathcal{L}$ has a branch i whose $N-1$ parallel branches $j_1, \dots, j_{N-1}$ have downstream node voltage measurements or estimates. Set $\ell = \{i, j_1, \dots, j_{N-1}\}$, remove ℓ from $\mathcal{L}$.
ELSEIF $\mathcal{L}$ has a branch i with downstream voltage measurements, select the branch and go to **4.**
ELSE Print warning that the circuit does not have sufficient measurement points and **STOP**.
ENDIF
3. IF ℓ has only one branch, estimate the impedance parameters of the branch in ℓ with the single-branch regression formulation (23)-(26) and go to **1.**
ELSE Estimate the impedance parameters of the branches in ℓ with the parallel branch regression formulation (30)-(33), estimate the voltages of the upstream node of the N parallel branches with (36), and go to **1.**
4. Estimate the parameters of the selected branch with (42)-(45) by forming a regression problem between the branch downstream node and the closest upstream node with voltage measurements. Go to **1.**
ENDIF

5.1.1 Estimating Series Branch Impedances

The step 4 of Algorithm 3 involves formulating a regression problem for a set of N series branches. This can be done based on the voltage drop over the series branches (22)

$$V_{Up} - V_{Down} = \sum_{i=1}^{N} (R_i I_{Ri} + X_i I_{Xi}) \tag{41}$$

where V_{Up} is the known voltage of the upstream node of the highest branch in the set and V_{Down} is the known voltage of the downstream node of the lowest branch in the set. The current components of branch i, I_{Ri} and I_{Xi}, can be calculated as a sum of the downstream branch currents of branch i. Consider T synchronous measurement samples $\boldsymbol{V}_{Up}, \boldsymbol{V}_{Down}, \boldsymbol{I}_{Ri}, \boldsymbol{I}_{Xi} \in \mathbb{R}^T, i \in \{1, \dots, N\}$ and define the response vector

$$\boldsymbol{y} = \boldsymbol{V}_{Up} - \boldsymbol{V}_{Down} \tag{42}$$

the measurement (design) matrix

$$\boldsymbol{\mathcal{X}} = [\boldsymbol{I}_{R1}, \boldsymbol{I}_{X1}, \boldsymbol{I}_{R2}, \boldsymbol{I}_{X2}, \dots, \boldsymbol{I}_{RN}, \boldsymbol{I}_{XN}] \tag{43}$$

and the parameter vector

$$\boldsymbol{\beta} = [R_1, X_1, \dots, R_N, X_N]^T \tag{44}$$

Then, the parameters $R_i, X_i, i \in \{1, \dots, N\}$ can be estimated from

$$\boldsymbol{y} = \boldsymbol{\mathcal{X}}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{45}$$

with one of the approaches introduced above. If the parameters of an upstream branch $i \in \{1, \dots, N\}$ are known (or estimated previously), branch i can be removed from the regression problem by doing the following steps.

1. Remove $\Delta\boldsymbol{V}_i$, the vector of voltage drops over branch i, from the response vector

$$\boldsymbol{y} = \boldsymbol{V}_{Up} - \boldsymbol{V}_{Down} - \Delta\boldsymbol{V}_i \tag{46}$$

2. Remove $\boldsymbol{I}_{Ri}$ and $\boldsymbol{I}_{Xi}$, the predictors of branch i, from the design matrix

$$\boldsymbol{\mathcal{X}} = \left[\boldsymbol{I}_{R1}, \boldsymbol{I}_{X1}, \boldsymbol{I}_{R2}, \boldsymbol{I}_{X2}, \dots, \boldsymbol{I}_{R(i-1)}, \boldsymbol{I}_{X(i-1)}, \dots, \boldsymbol{I}_{R(i+1)}, \boldsymbol{I}_{X(i+1)}, \dots, \boldsymbol{I}_{RN}, \boldsymbol{I}_{XN}\right] \tag{47}$$

3. Remove R_i and X_i, the coefficients of branch i, from the parameter vector

$$\boldsymbol{\beta} = [R_1, X_1, \dots, R_{i-1}, X_{i-1}, R_{i+1}, X_{i+1}, \dots, R_N, X_N]^T \tag{48}$$

If the parameters of multiple upstream branches are known, steps 1 – 3 are repeated for each branch with known parameters.

5.1.2 Three-Phase Test Circuit Results

First, Algorithm 3 operation is illustrated with different meters without voltage measurements in the secondary circuit 6 of the 66-node test circuit (Figure 33). Figure 59 shows both the true parameters and the estimated parameters when all loads have voltage measurements.

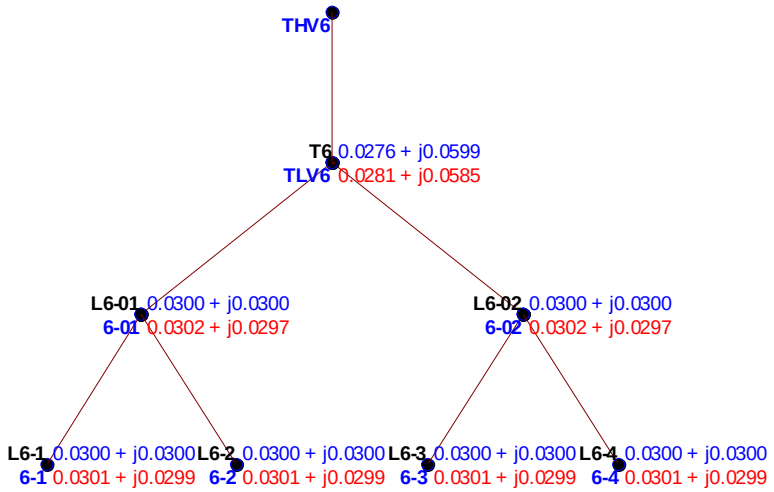

Figure 59. Estimated secondary circuit 6: node name (black bold), node upstream branch name (bold blue), branch true impedance (blue), and branch estimated impedance (red), branches whose parameters are not estimated are not shown

Figure 60 shows the estimated parameters when load 6-1 has no voltage measurements. In this case, branch L6-1 parameters are not estimated (and thus, are not shown in the figure), branch L6-2 parameters are estimated with the approach described in section 5.1.1 and the rest of the parameters are estimated normally utilizing voltage measurements.

Figure 61 shows the estimated parameters when loads 6-1 and 6-4 have no voltage measurements. In this case, branch L6-1 and L6-4 impedances are not estimated (and thus, are not shown in the figure), branch L6-2, L6-3 impedances are estimated with the approach from section 5.1.1, and the rest of the branches are estimated normally.

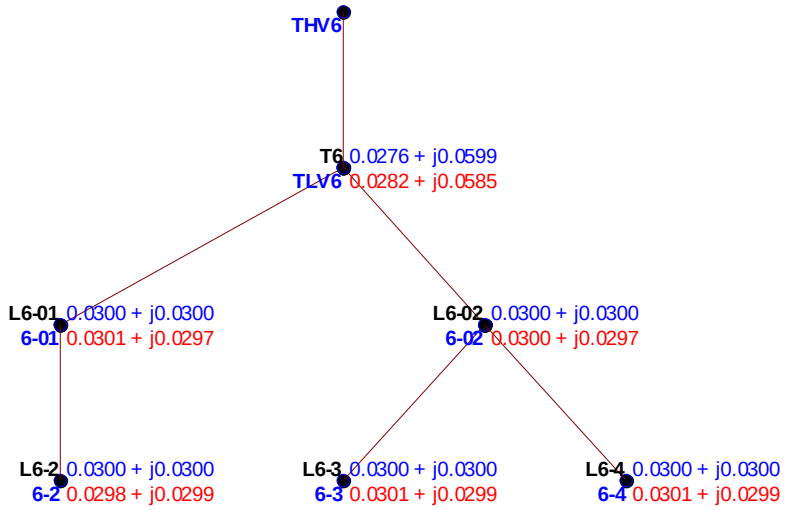

Figure 60. Estimated secondary circuit 6 parameters when load 6-1 has no voltage measurements: node name (black bold), node upstream branch name (bold blue), branch true impedance (blue), and branch estimated impedance (red), branches whose parameters are not estimated are not shown

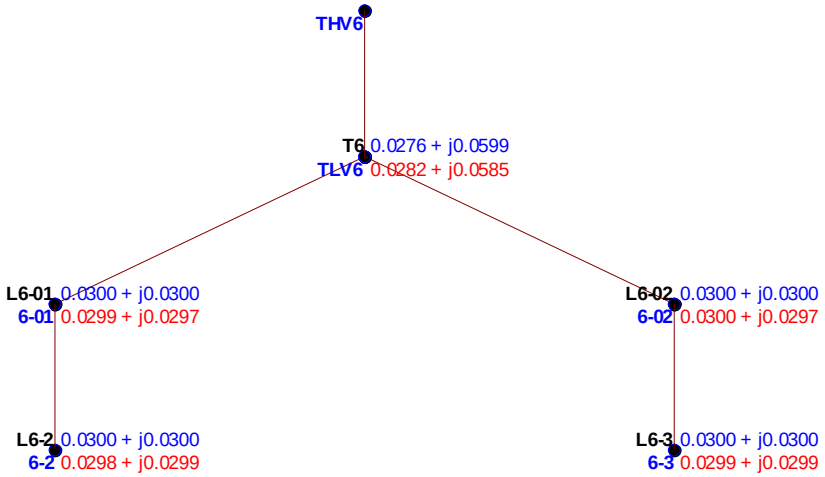

Figure 61. Estimated secondary circuit 6 parameters when loads 6-1 and 6-4 have no voltage measurements: node name (black bold), node upstream branch name (bold blue), branch true impedance (blue), and branch estimated impedance (red), branches whose parameters are not estimated are not shown

Finally, Figure 62 shows the estimated parameters when loads 6-1, 6-2, and 6-4 have no voltage measurements. In this case, branch L6_1, L6_2, L6_4, and L6_01 impedances are not estimated (and thus, not show in the figure), branch 6-3 impedance is

estimated with the approach in section 5.1.1 and the rest of the parameters are estimated normally.

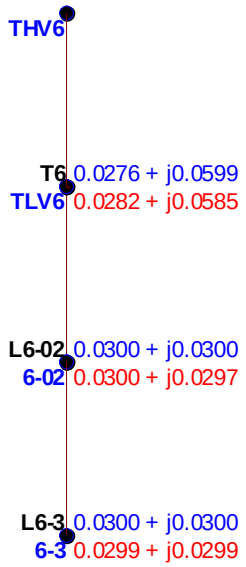

Figure 62. Estimated secondary circuit 6 parameters when loads 6-1, 6-2, and 6-4 have no voltage measurements: node name (black bold), node upstream branch name (bold blue), branch true impedance (blue), and branch estimated impedance (red), branches whose parameters are not estimated are not shown

The modified algorithm was validated on the 66-node test circuit by first removing voltage measurements from a given number of randomly selected meters in each secondary circuit and then estimating the parameters. This was repeated for 50 times for each secondary circuit. All the meters were assumed to have current measurements, i.e., no conversion from powers to currents with estimated/simulated voltages was necessary. The parameters were estimated with a set of 8760 samples of perfect measurements (no measurement error). Figure 63 and Figure 64 show the average (over the 50 repetitions) absolute relative errors of the estimated R and X, respectively. The more meters without voltage measurements a given secondary circuit has, the higher the errors of the estimated parameters become. In some secondary circuit topologies, the errors increase more than on others. This is clearly illustrated in Figure 65 and Figure 66 that show how much the average (over the 50 repetitions) absolute errors of the estimated R and X increase as the

number of meters with no voltage measurements increases. It is also interesting to observe that in some cases, e.g., branches “L5_2” and “T3”, the average parameter estimation error decreases when a meter is removed. This could potentially be explained by the particular load characteristics of the removed meter.

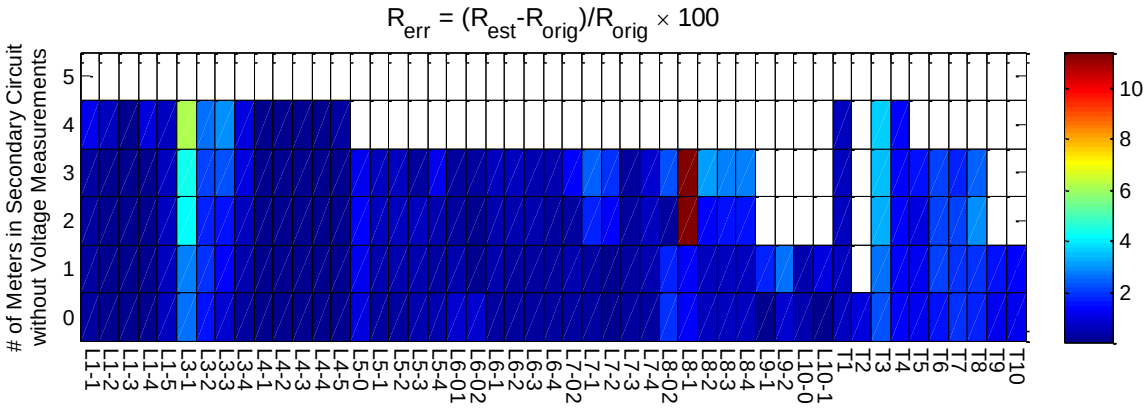

Figure 63. The average errors of the estimated R parameters over 50 repetitions where at each repetition a given number of randomly selected meters had no voltage measurements. In white areas, the parameter was not estimated in any of the repetition.

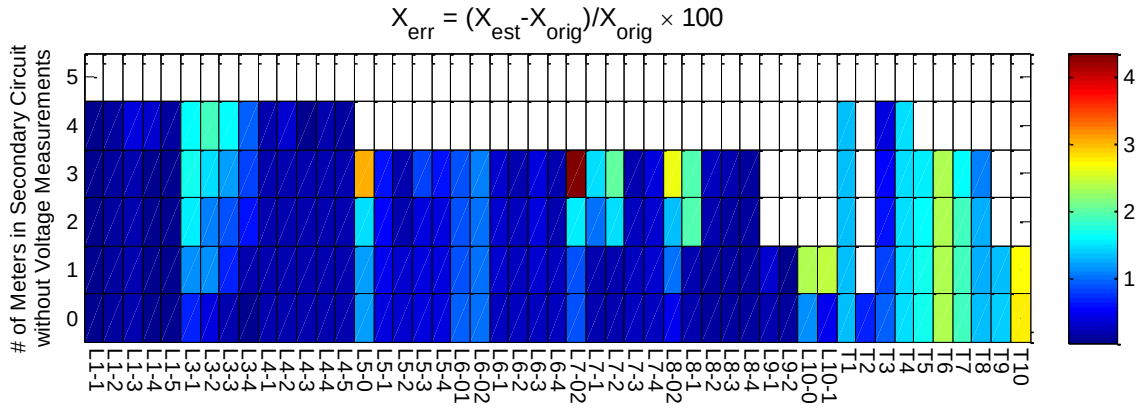

Figure 64. The average errors of the estimated X parameters over 50 repetitions where at each repetition a given number of randomly selected meters had no voltage measurements. In white areas, the parameter was not estimated in any of the repetition.

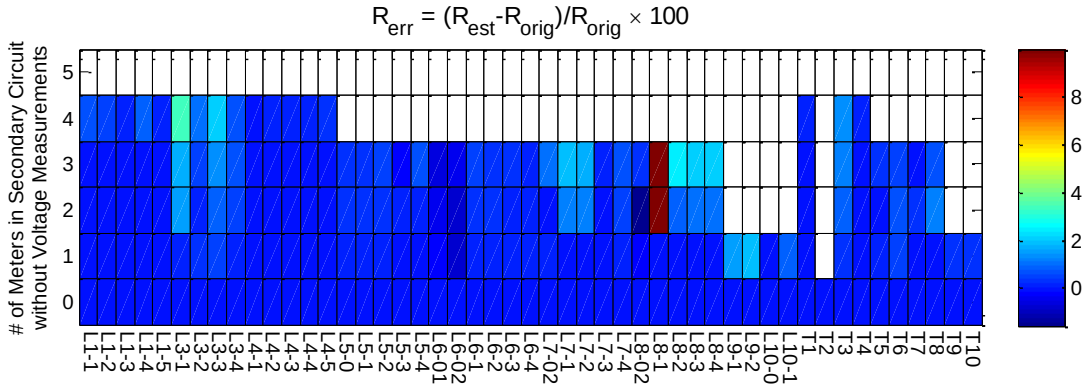

Figure 65. The impact of the number of meters with missing voltage measurements to the R estimation error (the difference of the results in Figure 63 compared to the case when all the meters have voltage measurements)

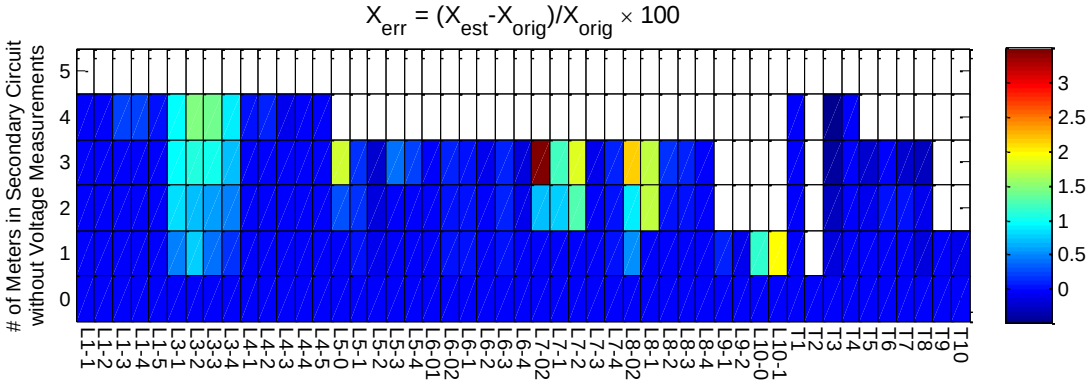

Figure 66. The impact of the number of meters with missing voltage measurements to the X estimation error (the difference of the results in Figure 64 to the case when all the meters have voltage measurements)

5.2 Simplified Algorithm with Limited PV Measurements

Ideally, secondary circuit parameters (and topologies) are estimated using a large set of synchronized historical voltage, active power, and reactive power measurement samples available from all the secondary circuit loads and distributed generation. In practice however, not all loads and DG units are metered and not all metered values are stored into a historical database. Moreover, some (especially older) meters may provide power (or current) measurements but no voltage measurements. A modified LRPE algorithm, that can handle some meters that do not transmit voltage measurements, is

shown in Section 5.1. This section addresses the currently common case when a utility does not have a dense network of smart meters (or other sensors) in the secondary circuits.

This section assumes that no (or very limited) AMI measurements are available but historical PV system measurements are available. Given this data, a simplified linear regression secondary circuit parameter estimation algorithm (SLRPE) is proposed to create simplified secondary circuit models shown in Figure 67 and to estimate their parameters. The objective is to improve the PV (or other sensor) voltage simulation accuracy. This section makes the following assumptions and simplifications.

1. Each secondary circuit (of interest) has one or more PV systems (or other sensor) measuring voltages and active and reactive powers V_{PV}, P_{PV}, Q_{PV} shown in blue in Figure 67. The discussion here focuses on one PV system in each secondary circuit but generalization to multiple PV systems is trivial.
2. Feeder total power measurements are available from SCADA.
3. The secondary circuit loads P_0, Q_0, P_1, Q_1 shown in green in Figure 67 are estimated with load allocation from the feeder total power SCADA measurements. In this dissertation, the load allocation is based on service transformer rating and the number of customers in each secondary circuit.
4. The feeder primary (medium-voltage) system is well-modeled including the service transformer connection.
5. The customer with a PV system is assumed to be connected to the service transformer secondary over a separate service line but the rest of the secondary circuit topology, component types, and component parameters are unknown.
6. Transformer primary side voltages referred to the low-voltage side, V_0, are estimated with time-series power flow simulation. Due to primary circuit modeling inconsistencies and the simplifications of load allocation, the simulated voltages may not be very accurate.

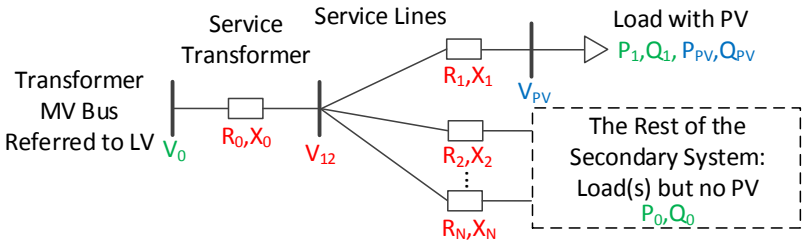

Figure 67. Simplified secondary circuit model with a PV system: available measurements are in blue, values that can be roughly estimated are in green, and unknown values and parameters are in red

Although Figure 67 does not represent all possible secondary circuit topologies, it is the best assumption given the limited available measurements and that the topology is unknown. Since the "rest of the secondary circuit" has no voltage measurements, it is not possible to estimate the impedances $R_2, X_2, \dots, R_N, X_N$ or any impedances in the "rest of the circuit". Thus, ignoring the losses in the "rest of the secondary circuit", the circuit in Figure 67 can be simplified to the circuit shown in Figure 68.

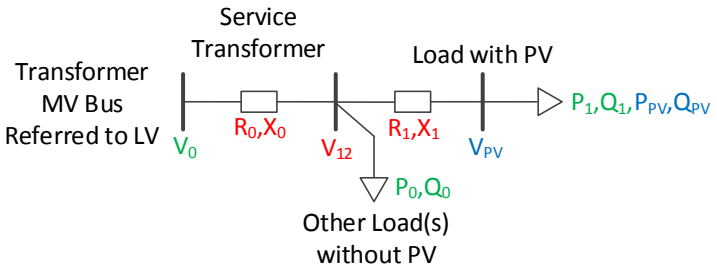

Figure 68. Simplified secondary circuit with a PV system: available measurements are in blue, values that can be roughly estimated are in green, and unknown values and parameters are in red

5.2.1 Method

5.2.1.1 Secondary Circuits with a Single PV System

The objective is to estimate the unknown parameters R_0, X_0, R_1, X_1 shown in red in Figure 68 by utilizing the available measurements V_{PV}, P_{PV}, Q_{PV} shown in blue in Figure 68 and the estimated measurements V_0, P_1, Q_1, P_0, Q_0 shown in green in Figure 68. This can

be achieved by utilizing M synchronous measurement samples in the linear regression formulation

$$\boldsymbol{V}_0 - \boldsymbol{V}_{PV} = R_0\boldsymbol{I}_{R0} + X_0\boldsymbol{I}_{X0} + R_1\boldsymbol{I}_{R1} + X_1\boldsymbol{I}_{X1} + \boldsymbol{\epsilon}. \tag{49}$$

The currents $\boldsymbol{I}_{R0}, X_0\boldsymbol{I}_{X0}, \boldsymbol{I}_{R1}, \boldsymbol{I}_{X1}$ are given by

$$\boldsymbol{I}_{R0} = (\delta_1\boldsymbol{P}_{SS} - \boldsymbol{P}_{PV})/\boldsymbol{V}_{PV} + \delta_0\boldsymbol{P}_{SS}/\boldsymbol{V}_{12} \tag{50}$$

$$\boldsymbol{I}_{X0} = (\delta_1\boldsymbol{Q}_{SS} - \boldsymbol{Q}_{PV})/\boldsymbol{V}_{PV} + \delta_0\boldsymbol{Q}_{SS}/\boldsymbol{V}_{12} \tag{51}$$

$$\boldsymbol{I}_{R1} = (\delta_1\boldsymbol{P}_{SS} - \boldsymbol{P}_{PV})/\boldsymbol{V}_{PV} \tag{52}$$

$$\boldsymbol{I}_{X1} = (\delta_1\boldsymbol{Q}_{SS} - \boldsymbol{Q}_{PV})/\boldsymbol{V}_{PV} \tag{53}$$

where $\boldsymbol{V}_{12}$ are the service transformer secondary voltages, $\boldsymbol{P}_{SS}, \boldsymbol{Q}_{SS}$ are the feeder total power measurements, δ_1 and δ_0 are the load allocation factors for the load at the PV and the other loads, respectively.

Often in practice however, it is not possible to use the linear regression formulation (49) directly and further simplifications are required. First, PV systems often operate at unity power factor ($\boldsymbol{Q}_{PV} = \boldsymbol{0}$). Second, $\boldsymbol{V}_{12}$ is unknown and cannot be estimated with the approach shown in section 4.2.2 because none of the other loads on the secondary circuit are assumed to have voltage measurements. For simplicity, the approximation of $\boldsymbol{V}_{12} \approx \boldsymbol{V}_{PV}$ can be used since the voltage drop over the service line is relatively small, and in practice, the errors resulting from generic load allocation without power measurements introduces much more error. Third, if no reliable measurements for the feeder total reactive power are available, $\boldsymbol{Q}_{SS}$ can be estimated from the feeder total active power measurements $\boldsymbol{P}_{SS}$ with a constant power factor (PF) with

$$\boldsymbol{Q}_{SS} = \sqrt{1/(PF)^2 - 1}\boldsymbol{P}_{SS} := \gamma\boldsymbol{P}_{SS}. \tag{54}$$

As a result of these three simplifications, the currents $\boldsymbol{I}_{R0}, X_0\boldsymbol{I}_{X0}, \boldsymbol{I}_{R1}, \boldsymbol{I}_{X1}$ in (50)-(53) are

$$\boldsymbol{I}_{R0} = \delta_1\boldsymbol{P}_{SS}/\boldsymbol{V}_{PV} - \boldsymbol{P}_{PV}/\boldsymbol{V}_{PV} + \delta_0\boldsymbol{P}_{SS}/\boldsymbol{V}_{PV} \tag{55}$$

$$\boldsymbol{I}_{X0} = \delta_1\gamma\boldsymbol{P}_{SS}/\boldsymbol{V}_{PV} + \delta_0\gamma\boldsymbol{P}_{SS}/\boldsymbol{V}_{PV} \tag{56}$$

$$\boldsymbol{I}_{R1} = \delta_1 \boldsymbol{P}_{SS}/\boldsymbol{V}_{PV} - \boldsymbol{P}_{PV}/\boldsymbol{V}_{PV} \tag{57}$$

$$\boldsymbol{I}_{X1} = \delta_1 \gamma \boldsymbol{P}_{SS}/\boldsymbol{V}_{PV}. \tag{58}$$

By defining

$$\boldsymbol{I}_{SS} = \boldsymbol{P}_{SS}/\boldsymbol{V}_{PV}, \tag{59}$$

and

$$\boldsymbol{I}_{PV} = \boldsymbol{P}_{PV}/\boldsymbol{V}_{PV}, \tag{60}$$

the currents $\boldsymbol{I}_{R0}, X_0\boldsymbol{I}_{X0}, \boldsymbol{I}_{R1}, \boldsymbol{I}_{X1}$ in (55)-(58) can be expressed with

$$[\boldsymbol{I}_{R0} \quad \boldsymbol{I}_{X0} \quad \boldsymbol{I}_{R1} \quad \boldsymbol{I}_{X1}] = [\boldsymbol{I}_{SS} \quad \boldsymbol{I}_{PV}] \begin{bmatrix} \delta_0+\delta_1 & \gamma(\delta_0+\delta_1) & \delta_1 & \gamma\delta_1 \\ -1 & 0 & -1 & 0 \end{bmatrix}, \tag{61}$$

clearly illustrating that there are only two linearly independent predictors. As a result, it is possible to estimate only two parameters with the available set of measurements. If the line per-unit-length resistance r and reactance x and the transformer X/R-ratio $(X/R)_0$ are assumed to be known, transformer resistance, R_0 (thus, effectively the transformer total impedance) and the line length, l_1, can be estimated with

$$\boldsymbol{V}_0 - \boldsymbol{V}_{PV} = R_0\boldsymbol{I}_0 + L_1\boldsymbol{I}_1 + \boldsymbol{\epsilon}, \tag{62}$$

where the currents are given by

$$\boldsymbol{I}_0 = \boldsymbol{I}_{R0} + (X/R)_0\boldsymbol{I}_{X0}, \tag{63}$$

$$\boldsymbol{I}_1 = r_1\boldsymbol{I}_{R1} + x_1\boldsymbol{I}_{X1}, \tag{64}$$

and $\boldsymbol{I}_{R0}, \boldsymbol{I}_{X0}, \boldsymbol{I}_{R1}, \boldsymbol{I}_{X1}$ are given in (61). Now, predictors $\boldsymbol{I}_0, \boldsymbol{I}_1$ are linearly independent provided that $\boldsymbol{I}_{SS}$ and $\boldsymbol{I}_{PV}$ are linearly independent, which occurs whenever $\boldsymbol{P}_{PV} \neq \boldsymbol{0}$. Once R_0 and L_1 have been estimated, the transformer reactance can be calculated with $X_0 = R_0(X/R)_0$ and the line impedances with $R_1 + jX_1 = L_1(r_1 + jx_1)$. As a result, the circuit in Figure 68 is further simplified to the circuit shown in Figure 69.

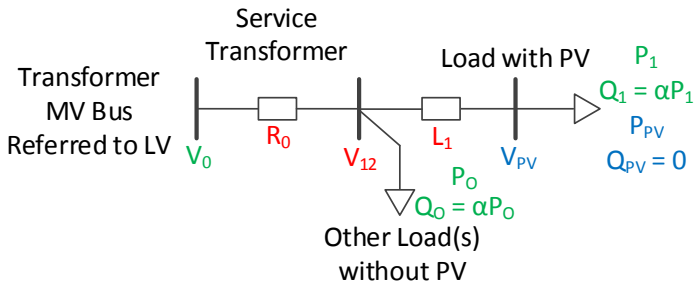

Figure 69. Further simplified secondary circuit with a PV system: available measurements are in blue, values that can be roughly estimated are in green, and unknown values and parameters are in red

5.2.1.1 Secondary Circuits with Multiple PV Systems

If the secondary circuit has $N > 1$ PV systems (model (62) is used for secondary circuits with only one PV system), the line parameters are first estimated with linear model

$$\boldsymbol{y} = \boldsymbol{\mathcal{X}}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{65}$$

where $\boldsymbol{\epsilon}$ represents the model and measurement error, the response variable $\boldsymbol{y}$ is given by

$$\boldsymbol{y} = \left[V_{PV1,1}, \dots, V_{PV1,M}, \dots, V_{PVN,1}, \dots, V_{PVN,M}\right]^{\mathrm{T}} \tag{66}$$

the unknown parameter vector is given by

$$\boldsymbol{\beta} = \left[V_{12,1}, \dots, V_{12,M}, R_1, X_1, \dots, R_N, X_N\right]^{\mathrm{T}} \tag{67}$$

and the design matrix $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{(MN)\times(M+2N)}$ is given by

$$\boldsymbol{\mathcal{X}} = \begin{bmatrix} \mathbf{I} & [-\boldsymbol{I}_{R,1} & -\boldsymbol{I}_{X,1}] & \cdots & \mathbf{0} \\ \vdots & & \vdots & \ddots & \vdots \\ \mathbf{I} & & \mathbf{0} & \cdots & [-\boldsymbol{I}_{R,N} \quad -\boldsymbol{I}_{X,N}] \end{bmatrix} \tag{68}$$

where $\mathbf{I} \in \mathbb{R}^{M\times M}$ are identity matrices, $\boldsymbol{I}_{R,i}, \boldsymbol{I}_{X,i} \in \mathbb{R}^{M\times 1}, i \in \{1, \dots, N\}$ are the branch current measurements, and the zero submatrices have suitable sizes. It should be noted that as long as $\boldsymbol{I}_{R,i} \neq \boldsymbol{I}_{X,i} \forall i \in \{1, \dots, N\}$ (non-zero PV generation), the columns of $\boldsymbol{\mathcal{X}}$ are linearly independent. After the line impedances have been estimated, the voltages of the upstream node of the branches $\boldsymbol{V}_{12}$ can be estimated as an average of the upstream node voltage estimates of the N branches with

$$\boldsymbol{V}_{12} = \frac{1}{N}\sum_{i=1}^{N}\left\|\boldsymbol{V}_i + (R_i + jX_i)(\boldsymbol{I}_{R,i} + j\boldsymbol{I}_{X,i})\right\| . \quad (69)$$

Then, the service transformer parameters are estimated with

$$\boldsymbol{V}_0 - \boldsymbol{V}_{12} = R\boldsymbol{I}_R + X\boldsymbol{I}_X + \boldsymbol{\epsilon} \quad (70)$$

As discussed in section 4.2.1, there are various ways to estimate the parameters from the linear models (62), (65), and (70). In this dissertation, the parameters were estimated with (ordinary least squares) linear regression. If linear regression resulted in negative (or too small) parameters, linearly constrained least squares formulation (29) was used.

5.2.1.2 Algorithm

The SLRPE algorithm is described in Algorithm 4.

Algorithm 4. SLRPE algorithm

Input: A list of the simplified secondary circuit branches, $\mathcal{L}$, with fields upstream and downstream node, number of parallel branches, branch current measurements $\boldsymbol{I}_R, \boldsymbol{I}_X$, branch node voltage measurements $\boldsymbol{V}$
Output: Secondary circuit branch estimation results including the estimated parameters R_{est}, X_{est} and their p-values R_{pval}, X_{pval}, and the regression model R^2- and MSE-values
IF the secondary circuit has only one PV system, estimate the line length L_1 and the transformer resistance R_0 with (62)-(64) and **STOP**.
ELSE
Estimate the line lengths $L_1, \dots, L_N$ with (65)-(68).
Estimate upstream node voltage of the branches $\boldsymbol{V}_{12}$ with (69).
Estimate the transformer parameters R_0 and X_0 with (70) and **STOP**.
ENDIF

5.2.2 Studied Utility Feeder Models

The SLRPE algorithm was applied to models of three real California utility feeders shown in Figure 70. This section introduces the feeder models and discusses some of the challenges related to the modeling of the feeder voltage regulating device operation. The feeder models, each of which consists of thousands of buses, lines, and loads, as well as

hundreds of transformers, were reduced using the approach shown in [109]. Specifically, the secondary circuits without PV systems were reduced to fixed-current loads at the service transformer primary. The secondary circuits with PV system(s) were converted to the simplified secondary circuit format illustrated in Figure 69 for one PV system. It should be noted that the original feeder models included only generic secondary circuit models consisting of a service transformer with typical parameters and a triplex cable feeding each load. Service line types were selected so that the lines had sufficient capacity to serve the loads. This is a common utility practice to model the secondary circuits. Table 8 lists the key feeder model characteristics.

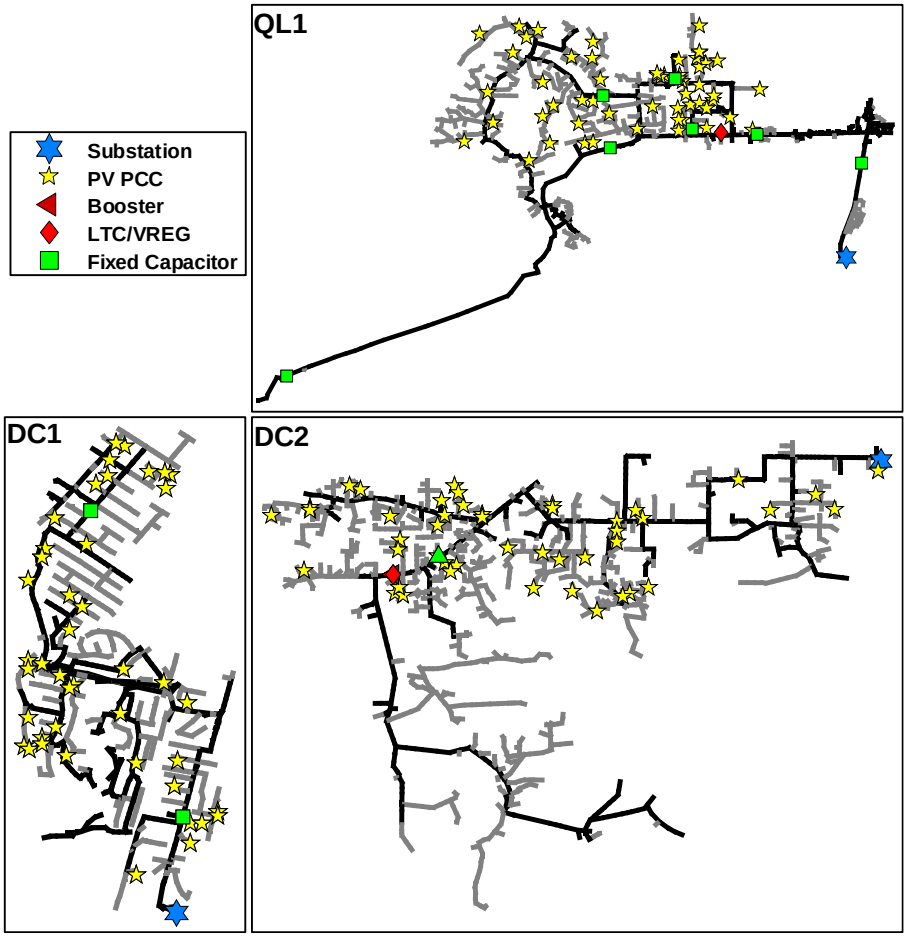

Figure 70. The topologies of the three full utility feeder models

Table 8. Full and reduced utility feeder model details

Feeder	QL1	DC1	DC2
Feeder Type	suburban	urban	rural
Voltage Level [kV]	20.78	12	12
# Customers	3500	3700	1200
Feeder Peak Load [MW]	18.63	8.08	3.6
Farthest 3-Phase Bus [km]	12.6	6.7	17.9
# PV Systems	44	36	31
LTC Set Point	120	123	121
# Capacitors – Control Mode	1-fixed 6-temperature	2-fixed	1-voltage
# Voltage Regulators	1	0	1
Available Reliable SCADA measurements	MW, MVAr phase currents	MW	MW

5.2.3 Method Validation for Feeder DC1

5.2.3.1 Full Secondary Circuit Models with Full AMI Data

First, perfect active power, reactive power, and voltage measurements were assumed to be available at all loads and the primary system was assumed to be perfectly modeled providing an accurate estimate of the primary side voltages of the service transformers. With these assumptions, the line and transformer impedance parameters of the 36 full (not simplified) secondary circuit models with PV systems were estimated with the LRPE algorithm using 744 measurement samples without measurement error. The average (longest) parameter estimation execution time for the 36 secondary circuits was 0.17 seconds (0.37 seconds). The average and maximum absolute relative errors of the estimated line and transformer parameters are shown in Table 9. Figure 71 shows the relative errors of the estimated line and transformer R and X parameters. Clearly, the LRPE algorithm estimates all the parameters with a very good accuracy.

Table 9. The average and maximum absolute relative errors of the estimated R and X

	$R_{err,avg}$ [%]	$X_{err,avg}$ [%]	$R_{err,max}$ [%]	$X_{err,max}$ [%]
Lines	0.004	0.027	0.029	0.309
Transformers	0.132	0.283	0.551	1.052

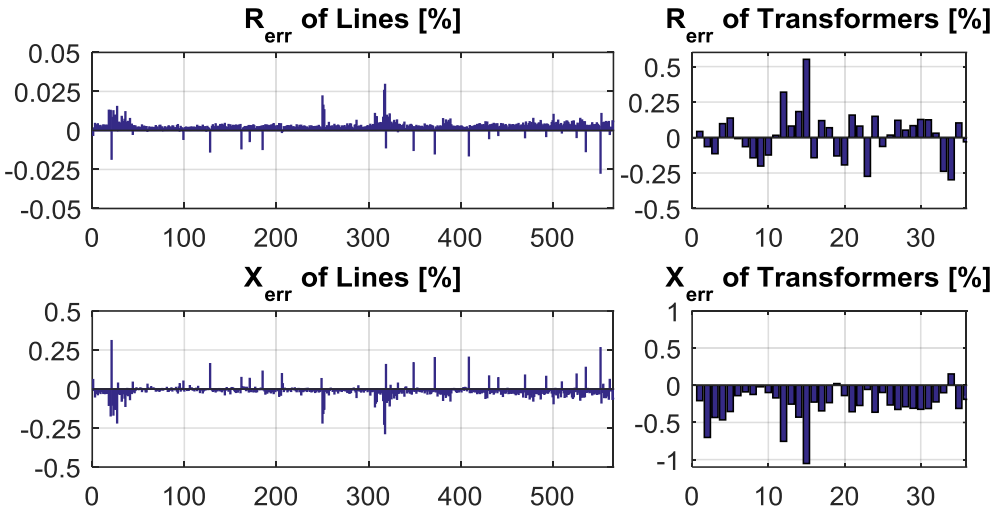

Figure 71. The relative errors of the full secondary circuit model parameters estimated with 744 samples of fully available AMI measurements without measurement noise

5.2.3.2 Simple Secondary Circuit Models with Full AMI Data

Next, the models of the 36 secondary circuits with PV systems were converted to the simple format shown in Figure 69 for one PV system where each secondary circuit had its original transformer, one generic service drop to the PV system (and the load at the PV system), and the rest of the secondary circuit loads were lumped at the service transformer secondary. All the loads were assigned the total feeder active power profile with a constant power factor. Then, the simple secondary circuit transformer resistance R_0 and the PV system service drop length L_1 were estimated with the SLRPE algorithm assuming that the load allocation was perfect, i.e., all loads follow the substation profile exactly. The transformer X/R-ratio and the line per-unit-length impedances r and x were also assumed to be perfectly known. The average (longest) parameter estimation execution time for the

36 secondary circuits was 0.07 seconds (0.24 seconds). The average (maximum) error of the estimated line length and transformer R parameters were 0.343% (2.347%) and 0.530% (1.394%), respectively. The relative errors of the line lengths and transformer resistances estimated with 744 samples without measurement error are shown in Figure 72.

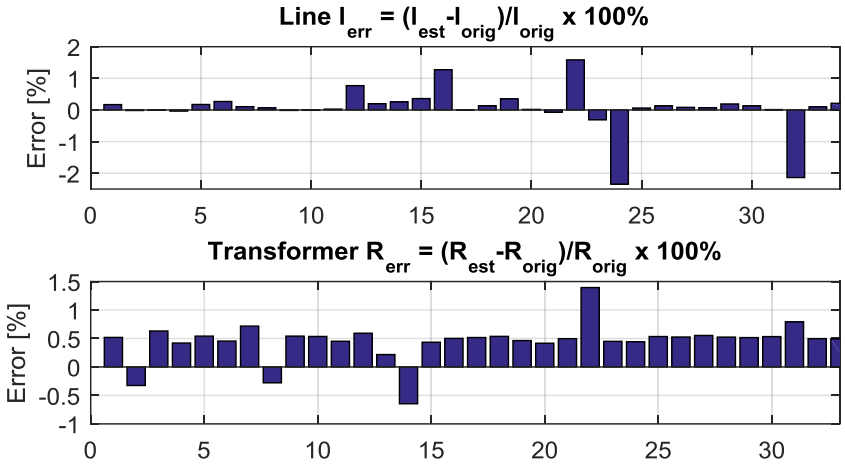

Figure 72. The relative errors of the simple secondary circuit parameters estimated with 744 samples without measurement noise. Simple secondary circuits with loads modeled through imperfect load allocation.

The higher overall errors in Figure 72 compared to Figure 71 can be mainly explained by the smooth profile of the allocated loads and the ill-conditioned linear regression design matrices. The design matrix ill-conditioning is caused by a high correlation of $\boldsymbol{I}_0$ and $\boldsymbol{I}_1$, which occurs, e.g., when the PV current is small compared to the current of the load located at the PV.

Each secondary circuit of the feeder DC1 has only one PV system and thus, the parameters are estimated with the regression model (62), where the predictor terms $\boldsymbol{I}_0$ and $\boldsymbol{I}_1$ are linear combinations of the currents $\boldsymbol{I}_{ss}$ and $\boldsymbol{I}_{PV}$. The linear combination coefficients depend on the transformer X/R-ratio, line per-line length parameters r and x, circuit power factor, and the number of loads in the secondary circuit. Thus, for secondary circuits with similar transformer and line types and similar number of customers, the predictor terms $\boldsymbol{I}_0$

and $\boldsymbol{I}_1$ differ only by the current $\boldsymbol{I}_{PV}$. This may partially explain why several parameter estimation errors are similar.

All transformer R parameters are estimated with an error less than 1% except for one secondary circuit where the difference of PV generation and its load were too small compared to the load at the transformer secondary leading to almost perfectly collinear predictor terms $\boldsymbol{I}_0$ and $\boldsymbol{I}_1$. All line parameters are estimated with an error less than 1% except for four lines that similar to the transformers, have highly correlated predictor terms $\boldsymbol{I}_0$ and $\boldsymbol{I}_1$ due to small PV generation and load at the PV relative to the other loads at the transformer secondary. Overall, these results indicate the theoretical feasibility of the SLRPE algorithm given that accurate load profiles or measurements are available.

5.2.3.3 Utility Feeder Results with Limited PV Measurements

Finally, the simple secondary circuit transformer resistance R_0 and the PV system service drop length L_1 were estimated with the SLRPE algorithm assuming that the loads are modeled through feeder total active power profile allocated to the loads based on service transformer rating. The average (longest) parameter estimation execution time for the 36 secondary circuits was 0.20 seconds (1.43 seconds). The average (maximum) error of the estimated line lengths L_1 and transformer resistances R_0 were 45.94% (318.0%) and 60.63% (91.05%), respectively. The estimated line length and transformer resistances are shown in Figure 73. In order to force the parameters to remain positive, the linearly constrained least squares estimation (29) was utilized to estimate the parameters in several secondary circuits. These results show that such typical load allocation approach does not sufficiently capture the load characteristics in individual secondary circuits to be useful for distribution system parameter estimation.

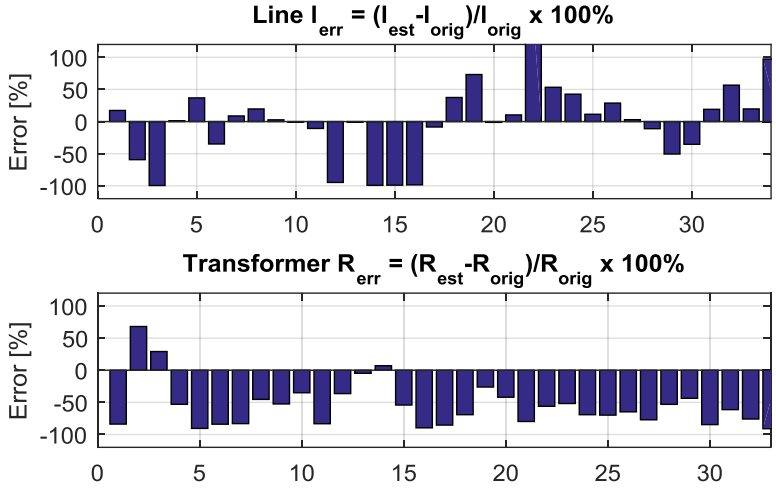

Figure 73. The relative errors of the simple secondary circuit parameters estimated with 744 samples without measurement noise. Simple secondary circuits with loads modeled through imperfect load allocation.

5.2.4 Results for the Studied Utility Feeders

This section presents the SLRPE algorithm results for the three utility feeder models introduced in section 5.2.2. Some of the feeder modeling challenges are discussed, validation results are listed, and the parameter estimation results are listed.

5.2.4.1 Substation LTC Modeling

Since the historical LTC primary voltages, secondary voltages, and taps were not available for the feeders, sub transmission was simply modeled as a constant Thevenin equivalent in all the three models. Moreover, the LTC voltage control set point were not known with a high confidence for any of the feeders. The LTC set points were selected to provide consistent results regarding positive average secondary circuit voltage drops with the smallest average differences between simulated PV voltages and measured PV voltages. Without substation voltage measurements, it was not possible to verify the accuracy of the simulated LTC medium-voltages.

Figure 74 shows feeder DC2 mean absolute errors and mean bias errors between the measured voltages and the voltages simulated with the original and the estimated

parameters with the selected set points of LTC being set at 121 Volts and the selected capacitor being switched off. The average (over the PV systems) mean absolute (bias) errors for the PV voltages simulated with the original and estimated parameters were 2.073 Volts (0.595 Volts) and 1.739 Volts (0.176 Volts), respectively. Similar plots were generated for different LTC set point and capacitor state comparison.

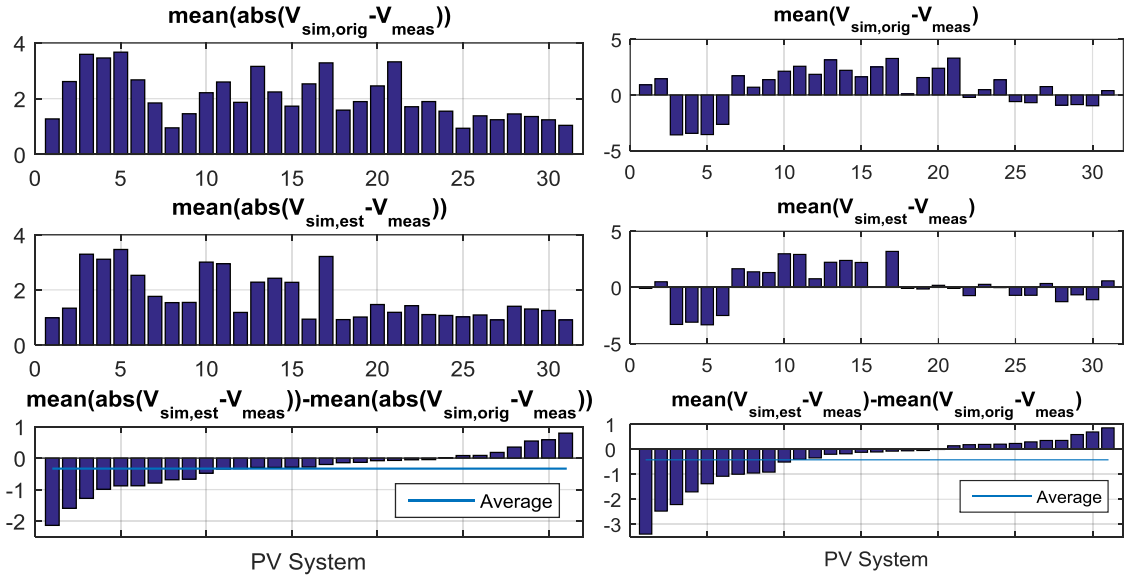

Figure 74. Feeder DC2 mean absolute errors (left) and mean bias errors (right) between the measured PV voltages and the PV voltages simulated with the original secondary circuit parameters (top) and the estimated parameters (middle). The lowest two plots show the error differences between the original and the estimated parameters sorted in ascending order.

5.2.4.2 Switched-Capacitor Bank Modeling

Feeders QL1 and DC2 have controlled capacitors that have a significant impact on the feeder voltage profile. Since the historical capacitor states were not available, it was necessary to estimate the capacitor states. Feeder QL1 capacitor states were estimated with the available SCADA feeder reactive power measurements. First, the total feeder reactive power load $Q_{load,tot}$ was estimated with load allocation from the feeder active power measurements using a constant power factor. Then, the total capacitor generated reactive power $Q_{cap,tot}$ was estimated by subtracting $Q_{load,tot}$ from the measured feeder reactive

power consumption $Q_{feeder,meas}$. Reactive power losses were neglected. When simulating the feeder voltages, the capacitors were turned on until simulated capacitor reactive power generation was close to the estimated capacitor reactive power generation. The capacitors were switched in a priority switching order, which was determined based on the capacitor temperature control set points. The simulated and estimated capacitor generated reactive power in May 2015 is shown in Figure 75. This capacitor state modeling scheme assures that the capacitor generated reactive power is approximately correct. However, given the limited available data, it is not possible to verify the true capacitor states.

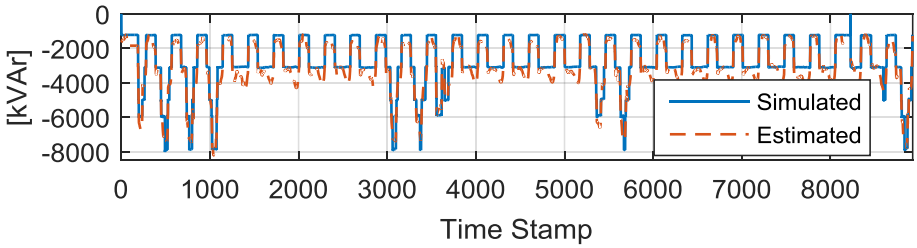

Figure 75. Feeder QL1 simulated and estimated capacitor generated reactive power in May 2015

Feeder DC2 has a voltage-controlled capacitor, whose set-point was not confidently known and had to be estimated. Since the feeder reactive power measurements were not reliable, the feeder capacitor state was estimated utilizing the PV voltage measurements. Based on the feeder model, the capacitor has a significant impact on the PV voltages as shown in Figure 76 on two right-most plots. The average absolute change in the simulated PV voltages was 7.45 volts when the capacitor switched on. Similar (but opposite) changes were observed in the simulated PV voltages when the capacitor was switched off. However, no significant changes were observed in the measured PV voltages as shown next.

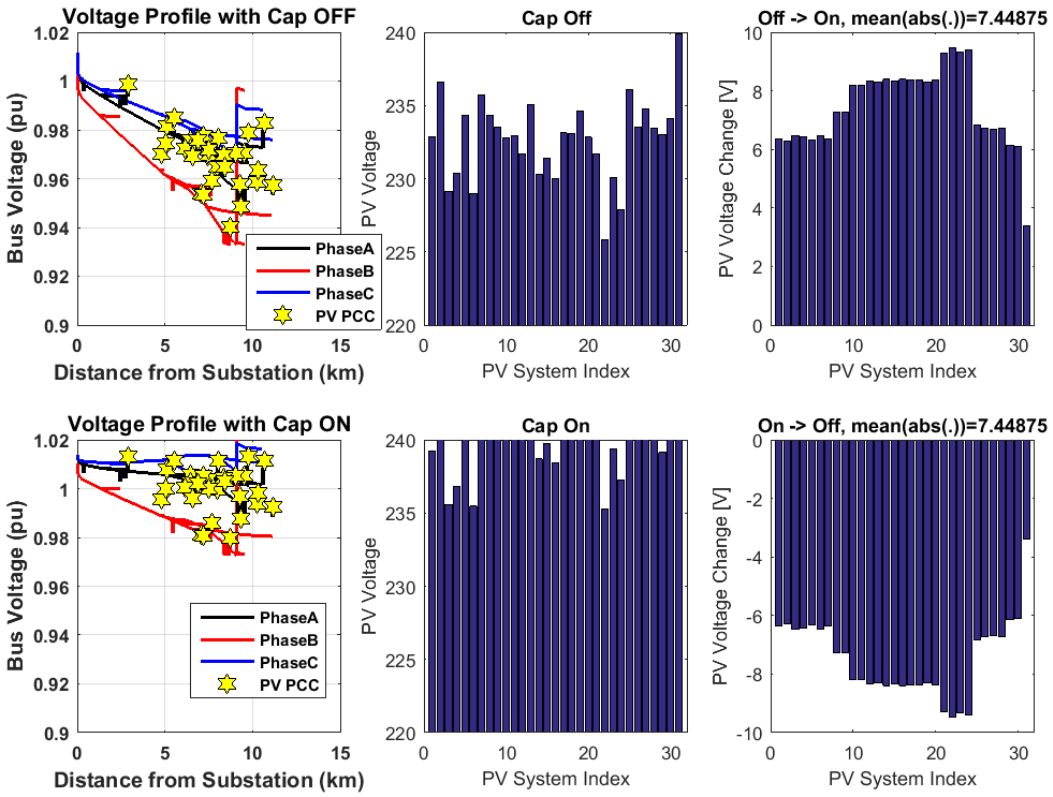

Figure 76: Feeder DC2 voltage profiles (left), simulated PV voltages (middle), and the simulated PV voltage changes when the capacitor is off (top) and on (bottom)

Figure 77 on the top shows the change in the measured PV voltages from one sample to the next. Figure 77 on the middle and on the bottom show the average (over the PV systems) absolute values and average (over the PV systems) values of the changes in the measured PV voltages. Although, the measured voltages of some PV systems have changes in the order of 5-7 volts, these changes are not consistent over the PV systems as expected for capacitor state changes based on the feeder model as shown in Figure 76. This can be verified from Figure 78, which shows the histograms of average absolute and average changes (from one sample to the next) in the measured PV voltages in feeder DC2. Since there were no changes even close to the level of 7.45 volts, it was assumed that the capacitor state did not change during the entire month. It should be noted that no PV measurements were available during the night time and it is possible that the capacitor state

changed during the nigh-time. However, without reliable feeder reactive power measurements or PV night-time voltage measurements, it was impossible to determine if capacitor state changes occurred during the night-time.

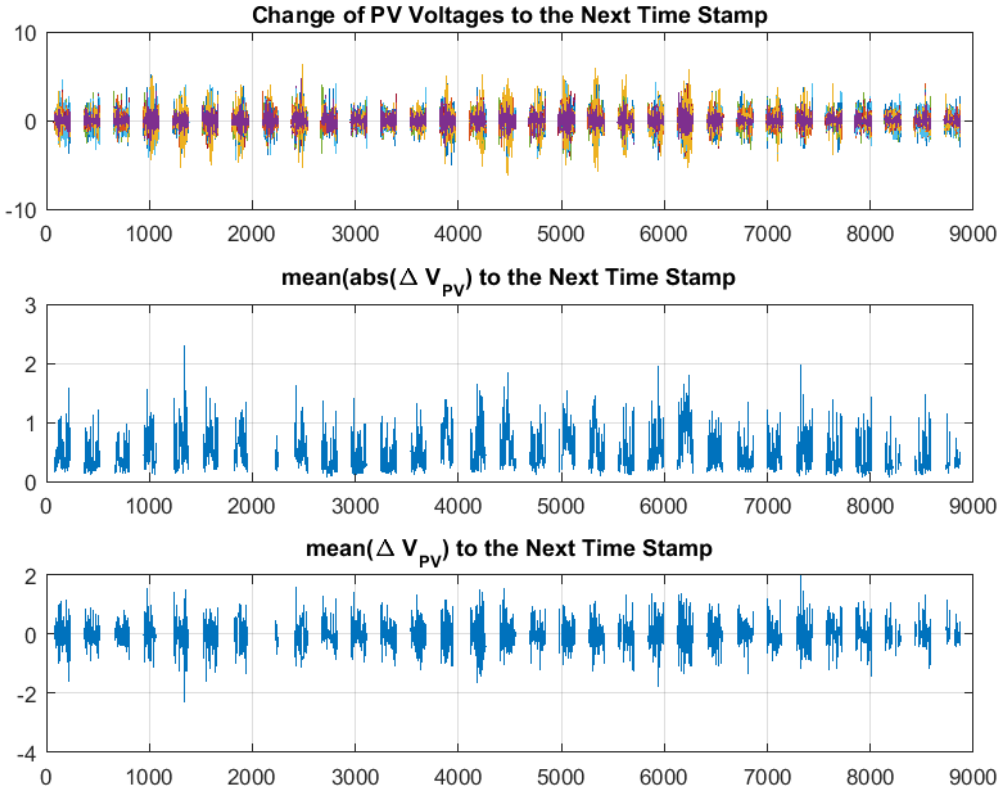

Figure 77: Changes in the measured PV voltages between samples (top), mean absolute changes (middle), and mean changes (bottom)

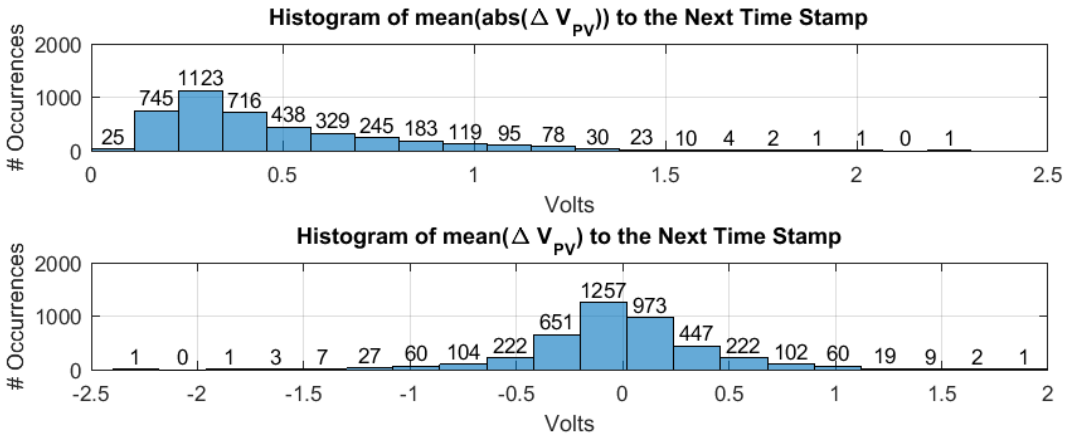

Figure 78: Histograms of mean absolute (top) and mean (bottom) changes in the measured PV voltages

5.2.4.3 Method Validation with Simulated Voltages

The simplified secondary circuit parameters of the three feeder models were estimated with 8928 samples (one month of 5-min samples) of simulated PV and service transformer voltages. The PV voltages and service transformer medium-voltages were simulated with time-series power flow with measured PV powers and load powers modeled with load allocation. The parameter estimation results with simulated voltages are summarized in Table 10. All parameters were estimated with linear regression very close to the original parameters. The linearly constrained least squares algorithm was not needed to force the parameters to be positive. The minimum R-squared value of the parameter estimation linear regression problems was above 0.9999 indicating that all the linear regression models provided an excellent fit to the data.

Table 10. Parameter estimation accuracy with the simulated PV voltages and service transformer medium-voltages

Feeder	Average Absolute Error of		Maximum Absolute Error of	
	$\frac{(R_{est} - R_{orig})}{R_{orig}}$	$\frac{(X_{est} - X_{orig})}{X_{orig}}$	$\frac{(R_{est} - R_{orig})}{R_{orig}}$	$\frac{(X_{est} - X_{orig})}{X_{orig}}$
QL1	0.26	0.25	0.96	0.96
DC1	0.30	0.33	3.57	3.57
DC2	0.33	0.34	1.25	1.25

5.2.4.4 Parameter Estimation With PV Voltage Measurements

Next, the simplified secondary circuit parameters of the three feeders were estimated with 8928 samples (one month of 5-min samples) of actual PV voltage measurements. The transformer medium-voltages were simulated with time series power flow utilizing the measured PV generation and load-allocated loads. Figure 79 shows absolute average differences between the measured and the simulated PV voltages. Simulating the PV voltages with the estimated parameters (as opposed to the original generic feeder parameters) effectively reduces the average absolute voltage simulation

errors on average by 0.57 Volts (19.3% reduction), 1.64 Volts (71.5% reduction), and 0.40 Volts (22.5% reduction), for feeders QL1, DC1, and DC2, respectively.

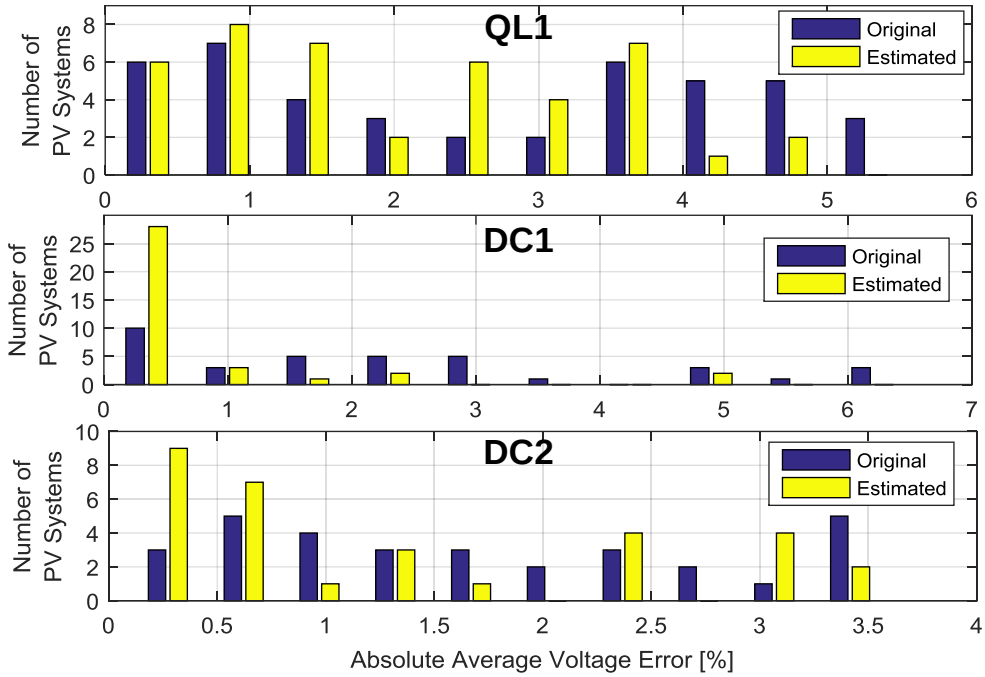

Figure 79. Feeder QL1 (top), DC1 (middle), and DC2 (bottom) absolute average voltage error for each PV system simulated with the original (in blue) and estimated (in yellow) parameters

5.3 Discussion

Algorithm 3 presented in this chapter extends the scope of the LRPE algorithm presented in Chapter 4 for handling meters that do not report voltage measurements. Although the Algorithm 3 has good accuracy when few meters do not report voltage measurements, any meter without voltage measurements reduces the accuracy and observability of the (secondary circuit) parameter estimation and thus, it is desirable to have high-quality voltage measurements from all smart meters. Moreover, ideally all meters would report current measurements I_R and I_X (or current magnitude and power factor for calculating I_R and I_X), since this would allow to avoid errors resulting from converting powers to currents with estimated or simulated voltages.

The SLRPE algorithm presented in this chapter allows generating simplified secondary circuit models when only limited measurements are available from PV systems or other sensors. This algorithm can be utilized, e.g., for improving the voltage simulation accuracy at the PV systems. The SLRPE algorithm results shown in this chapter demonstrate that PV measurement data can be effectively utilized to improve feeder voltage simulation accuracy by validating feeder voltage regulating device modeling and by performing secondary circuit parameter estimation. Next, some of the key findings from parameter estimation on the three utility feeder models are highlighted.

Figure 80 shows the boxplots of the PV voltage simulation errors for the feeder DC2. Parameter estimation is unable to reduce the variance in the voltage simulation errors, which results from the modeling inaccuracies and simplifications. In some cases, these inaccuracies and simplifications resulted in unrealistic parameters such as extremely long service line lengths as shown in Figure 81. The two main sources of error for the analyzed feeders seemed to be inaccuracies in the medium-voltage level modeling and modeling loads through load allocation.

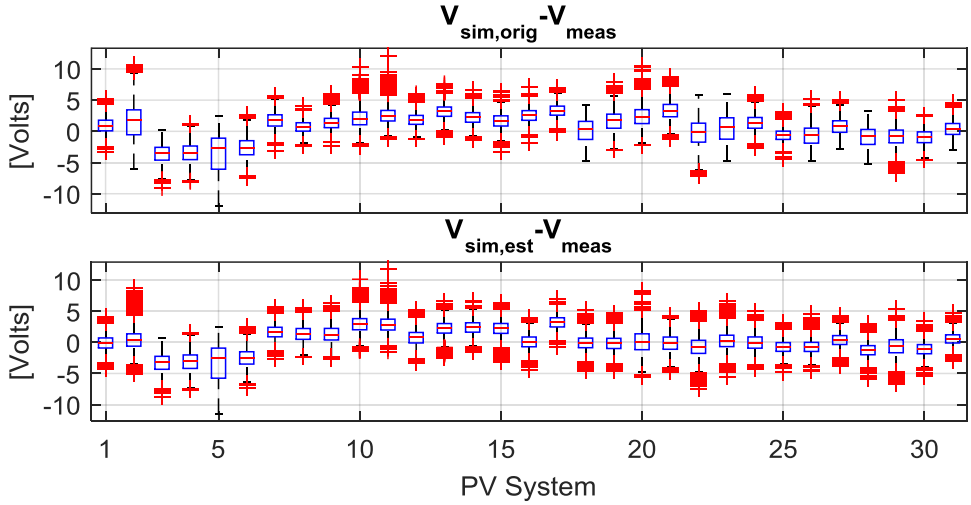

Figure 80. Feeder DC2 error of PV bus voltages simulated with the original (top) and estimated (bottom) parameters

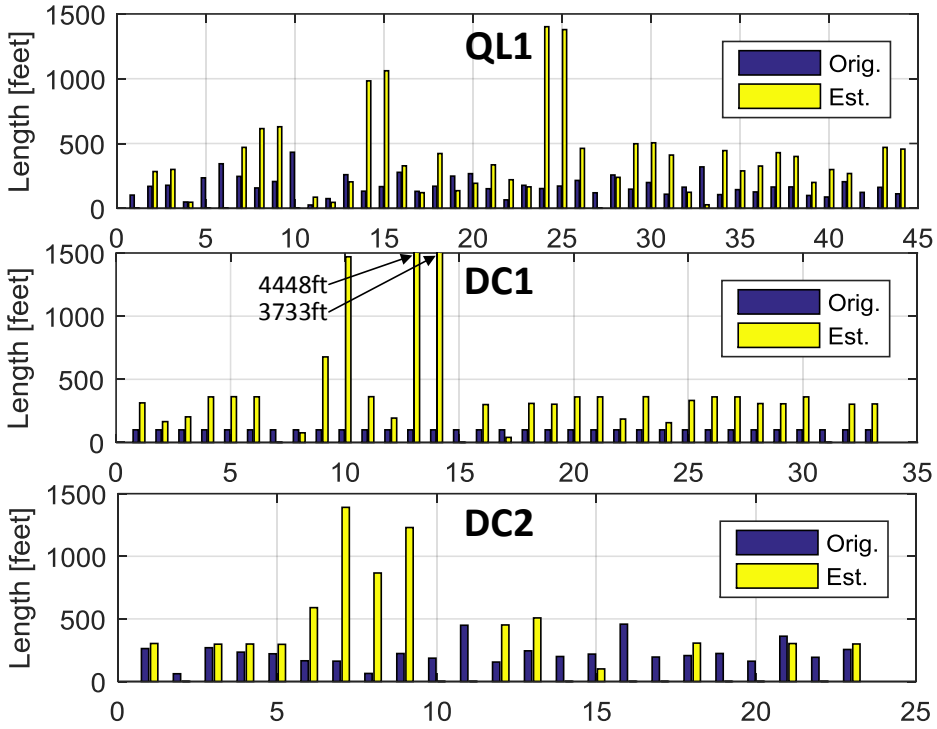

Figure 81. Original (in blue) and estimated (in yellow) service line lengths for the feeders QL1 (top), DC1 (middle), and DC2 (bottom)

As shown in section 4.6, parameter estimation is much more sensitive to voltage measurement error than power (or current) measurement error. Therefore, when utilizing simulated voltages to estimate, e.g., service transformer impedances, the voltages must be accurately simulated. For the analyzed utility feeder models, it turned out to be very challenging to identify the capacitor and LTC states with limited data. In order to reach good accuracy in parameter estimation regression models that utilize simulated voltages, it is necessary to correctly model the voltage-regulating device operation. Otherwise, one can observe negative or unrealistically high simulated voltage drops over the secondary circuits that the parameter estimation is trying to model by adjusting the impedance parameters. Utilities usually monitor the tap position of the substation LTC or in some cases, the LTC tap position is estimated by the transmission level state estimator. Ideally, the states of all the voltage regulating devices would be monitored and stored into SCADA. However,

before this becomes a common practice, future work should analyze including the voltage regulating device states (e.g. LTC tap position, capacitor on/off state) or parameters (e.g. LTC voltage control parameters) in parameter estimation.

Modeling secondary circuit loads through load allocation, which is based on substation SCADA measurements and service transformer rating (or similar metric), significantly simplifies the impacts on the secondary circuit level [24], [48], [110]. In general, using load allocation tends to underestimate the voltage drops and losses in the secondary circuits. In the simplified secondary circuit parameter estimation with one PV system (62)-(64), the response variable is $\boldsymbol{V}_0 - \boldsymbol{V}_{PV}$, where $\boldsymbol{V}_0$ is the simulated (with allocated loads) service transformer medium-voltages referred to the secondary and $\boldsymbol{V}_{PV}$ are the measured PV voltages. Due to the load allocation, $\boldsymbol{V}_0$ tends to have a very smooth profile over time and tends to overestimate the true voltages since the load allocation underestimates the secondary circuit voltage drops. On the other hand, $\boldsymbol{V}_{PV}$ can have a highly varying profile due to the characteristics of the PV generation and the load at the PV. As a result, the estimated voltage drop $\boldsymbol{V}_0 - \boldsymbol{V}_{PV}$ tends to vary more than it does in reality. Similarly, due to load allocation modeled loads, the predictor term $\boldsymbol{I}_0$ tends to have a relatively smooth profile compared to the predictor term $\boldsymbol{I}_1$, which varies much more over time due to the varying PV generation. Since the response variable $\boldsymbol{V}_0 - \boldsymbol{V}_{PV}$ is better correlated with the $\boldsymbol{I}_1$ than with the predictor variable $\boldsymbol{I}_0$, the line length L_1 tends to be overestimated and the transformer resistance R_0 tends to be underestimated.

Especially when the PV generation is small compared to the load consumption at the PV, the predictors $\boldsymbol{I}_0$ and $\boldsymbol{I}_1$, which are calculated based on allocated loads and the PV generation, are unable to explain almost any of the variation in the response variable (calculated as a difference of simulated and PV measured voltage). Figure 82 shows an example of such a regression problem where the response variable is highly uncorrelated with the predictor variables (which are normalized in the figure to allow displaying in the

same plot display). As a result, the regression model residuals (the model errors) follow the response variable almost perfectly.

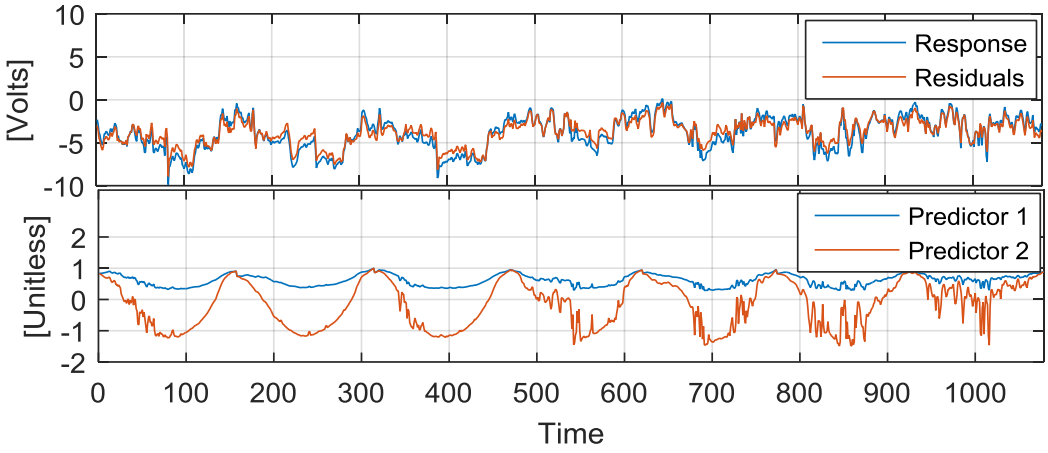

Figure 82. An example of a poor linear regression fit: response and residuals (top) and normalized predictors (bottom)

Overall, these results demonstrate that generic load allocation from substation data cannot be used for secondary circuit parameter estimation and that all injection points in the secondary network should have meters, such as AMI, in order to estimate the impedances. When all the secondary circuit injection measurements and some but not all voltage measurements are available, the secondary circuit model parameters can be estimated with one of the approaches shown in this chapter. If the topology is known (unknown), the approach shown in section 5.1 (section 5.2) should be utilized. When all the injection and voltage measurements are available but the secondary circuit topology is unknown, the topology (and the parameters) can be estimated with the method presented in Chapter 6.

CHAPTER 6. DISTRIBUTION SECONDARY CIRCUIT TOPOLOGY ESTIMATION

The parameter estimation methods presented in chapters 4 and 5 required the secondary circuit topology to be known. However, in many cases, the secondary circuit topologies are not known with a high confidence and thus, the topology needs to be estimated utilizing the measurement data. This chapter presents a method for jointly estimating the secondary circuit topology and the impedance parameters. First, section 6.1 introduces the joint secondary circuit topology and parameter estimation problem (DSTE). Section 6.2 demonstrates that for most secondary circuits, it is impossible to perform an exhaustive search of all topologies. Section 6.3 presents an algorithm for joint topology and parameter estimation. Section 6.4 presents an algorithm for validating an estimated topology in test cases where the true topology is known. Section 6.5 presents topology estimation results for a test circuit and for the secondary circuits of a Georgia Tech feeder. Finally, section 6.7 discusses the path towards practical implementation of distribution system topology and parameter estimation. Most of the work presented in this chapter has been published in [98], [99], [111].

6.1 Problem Formulation

The objective of the DSTE problem is to simultaneously identify topology (component connectivity) and component series impedance parameters of a given secondary circuit. We consider secondary circuits that mathematically are rooted trees, i.e., connected, directed graphs without cycles that have a designated root node (the secondary circuit transformer low-voltage node). Secondary circuit trees may be multifurcating, i.e., a given node may have more than two child nodes. Because the topology is unknown, the total number of nodes is also unknown.

The proposed LRTE algorithm requires that each secondary circuit tree leaf node has a smart meter or DER sensor measuring voltage and active and reactive power (or current and power factor) shown in blue in Figure 83. The voltages and injections at some of the internal nodes of the tree may also be metered. Utilizing this data, the objective is to identify the circuit topology and the component series impedance parameters shown in red in Figure 83. It should be emphasized that in this chapter the secondary circuit topology and component parameters are assumed to be completely unknown, except for the fact that it must be a tree.

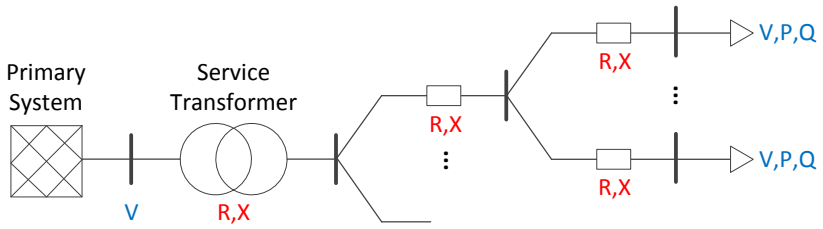

Figure 83. Secondary circuit topology and parameter estimation problem

6.2 Infeasibility of Exhaustive Topology Search

Theoretically, the topology of a secondary circuit can be estimated by performing an exhaustive search of all possible topologies. This could be achieved by first estimating the parameters for all possible topologies with the approach in Chapter 4 and then selecting the topology that results in the best accuracy, e.g., in terms of the mean squared error (MSE) of the simulated voltages compared to the measured voltages. This approach can quickly become infeasible for a secondary circuit with several meters, due to the number of possible topologies to consider.

As follows, the secondary circuit nodes that have a meter are referred to as labeled nodes and the order of unlabeled nodes within the secondary circuit tree is considered irrelevant for the purposes of this chapter. The number of possible secondary circuit topologies is given by the number of rooted potentially multifurcating trees with N labelled

nodes some of which may be internal nodes (allowing some meters to be at the internal nodes of the secondary circuit). The number of such trees is of interest in evolutionary biology and has been calculated with a recursive relation in [112]. Table 11 lists the number of such trees for $N \in \{1, \dots, 10\}$ labeled nodes both not allowing and allowing internal nodes to be labeled. The number of possible topologies is larger if labeled internal nodes are allowed. Clearly, even if labeled internal nodes are not allowed (all meters at the leaf nodes of the secondary circuit tree), it becomes impractical to evaluate all the alternative topologies with 5-7 meters and is practically infeasible with 8 or more meters. Thus, an exhaustive comparison of all possible topologies would be a computationally demanding task. The next subsection presents a computationally efficient greedy-type joint parameter and topology estimation approach.

Table 11. The number of rooted trees with N labelled nodes, allowing multifurcations, and either not allowing or allowing some of the internal nodes to be labeled [112]

Number of Labelled Leaf Nodes	Not Allowing Labeled Interior Nodes	Allowing Labelled Interior Nodes
1	1	1
2	1	3
3	4	22
4	26	262
5	236	4,336
6	2,752	91,984
7	39,208	2,381,408
8	660,032	72,800,928
9	12,818,912	2,566,606,784
10	282,137,824	102,515,201,984

6.3 Joint Parameter and Topology Estimation Algorithm

This section introduces linear regression topology and parameter estimation algorithm (LRTE) for jointly estimating secondary circuit topology and parameters. Similarly to the LRPE and SLRPE algorithms, the LRTE algorithm utilizes the linearized voltage drop approximation (22) and processes one secondary circuit at a time. For each

secondary circuit, the algorithm is initialized with the list of the meters in the secondary circuit and an empty list of mock circuit branches. For each meter pair at each iteration the algorithm solves a linear regression problem for the parallel circuit type (on the left in Figure 84)

$$\boldsymbol{V}_1 - \boldsymbol{V}_2 = R_1\boldsymbol{I}_{R1} + X_1\boldsymbol{I}_{X1} + R_2\boldsymbol{I}_{R2} + X_2\boldsymbol{I}_{X2} + \boldsymbol{\epsilon} \tag{71}$$

and a linear regression problem for the series circuit type (on the right in Figure 84)

$$\boldsymbol{V}_1 - \boldsymbol{V}_2 = R\boldsymbol{I}_R + X\boldsymbol{I}_X + \boldsymbol{\epsilon} \tag{72}$$

The order of meters 1 and 2 is irrelevant in regression model (71). If the secondary circuit does not have distributed generation causing reverse power flows, the regression model (72) is solved only for the meter order with positive average voltage drop $\sum_{t=1}^{T}(V_{1,t} - V_{2,t}) > 0$. A wrong meter order simply results in negative estimated parameters. If the OLS estimator (28) results in unrealistic (negative or too large) parameters for the models (71) and (72), the CLS estimator (29) is used instead for that meter pair. The upstream node voltages of the parallel branch pair can be estimated with (36).

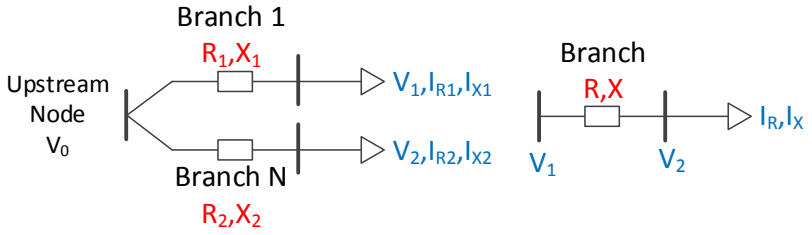

Figure 84. Two meters connected in parallel (left) and in series (right)

The complete LRTE algorithm is listed in Algorithm 5. The algorithm begins with a list of active meters equal to all the meters in the secondary circuit. When the list of active meters has only one meter left, Algorithm 5 stops and returns the mock circuit consisting of a list of mock circuit branches, $\mathcal{B}$, with the fields: from node names, to node names, and

impedances R and X. The mock circuit includes all the meters in the original secondary circuit.

Algorithm 5. LRTE algorithm

Input: Samples $\boldsymbol{V}, \boldsymbol{I}_R, \boldsymbol{I}_X$ from all the metered nodes
Output: Mock circuit consisting of a list of branches, $\mathcal{B}$ with the fields: from node names, to node names, estimated impedances R_{est} and X_{est}, their p-values R_{pval} and X_{pval}, and the regression model R^2- and $RMSE$-values.
1. Initialize the list of active meters, $\mathcal{L}$, as the list of all meters in the secondary circuit. Set, $\mathcal{B}$ empty.
2. IF $\mathcal{L}$ has only one meter, STOP.
3. FOR all the meter pairs in $\mathcal{L}$
a) Estimate the impedance parameters R_1, X_1, R_2, X_2 with the parallel meter pair linear regression formulation (71) using the OLS estimator (28).
b) For the meter pair order with positive average voltage drop, estimate the impedance parameters R, X with the series meter pair linear regression formulations (72) using the OLS estimator (28).
c) IF both 3. a) and 3. b) resulted in unrealistic parameter estimates or insignificant p-values
Set parts 3. a) and 3. b) RMSE values very large
Re-do parts 3. a) and 3. b) with the CLS estimator (29)
ENDFOR
4. Select the linear regression model with the best fit in terms of the smallest root mean squared value of the residuals. Denote the selected model M_{12} and the corresponding meters m_1 and m_2.
IF model M_{12} is of the parallel type (71)
5. Add two new branches with the impedance parameters obtained from model M_{12} to $\mathcal{B}$.
6. Add a new virtual upstream node with the sum of meter m_1 and m_2 currents and node voltages estimated with (36) to $\mathcal{L}$.
7. Remove meters m_1 and m_2 from $\mathcal{L}$.
ELSEIF model M_{12} is of the series type (72)
8. Add one new branch with the impedance parameters obtained from M_{12} to $\mathcal{B}$.
9. Add downstream meter m_2 currents to the upstream meter m_1 currents.
10. Remove meter m_2 from $\mathcal{L}$.
ENDIF
11. Go to Step **2.**

The joint topology and parameter estimation approach shown in [82] adds artificial close-to-zero-impedance branches in the common cases of three or more parallel meters

and two or more meters in series. The artificial branch creation is illustrated in Figure 85 for circuit subsections with four parallel branches (top plots) and with three series meters (bottom plots). If the impedances of these artificial branches are small, they can be removed from the circuit by, e.g., removing all branches with impedance below a selected threshold. However, it was observed to be challenging to set the impedance threshold to correctly remove the artificial branches while still preserving all the true branches. The higher the measurement error level is, the harder it seems to be to filter the artificial branches away. By utilizing the two circuit types shown in Figure 84, Algorithm 5 does not create any artificial close-to-zero-impedance branches provided that at each iteration, the algorithm selects the correct meter pair and regression model type.

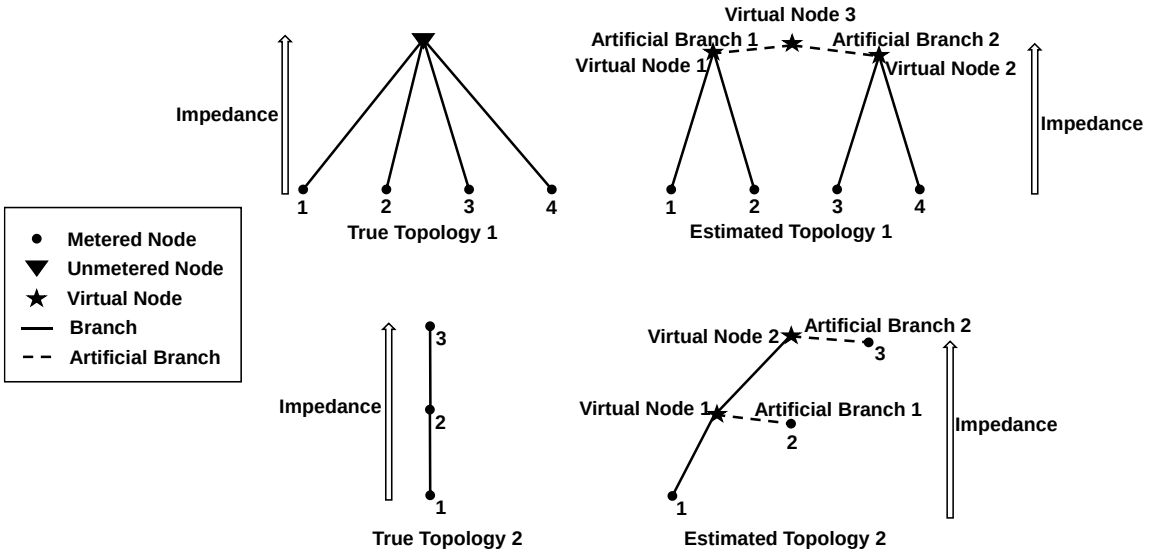

Figure 85. Example of artificial close-to-zero impedance branch creation: true (top left) and estimated (top right) circuit subsection topology with four parallel branches (top left) and true (bottom left) and estimated (bottom right) circuit subsection with three series meters.

6.4 Topology Estimation Result Validation

To allow avoiding cumbersome manual validations, this section presents an automated method for comparing an estimated circuit topology with a true circuit topology. The presented method is intended for topology estimation algorithm validation in test

circuits, where the true topology is known. First, the requirements for two secondary circuit topologies to have electrically identical topologies are listed. Then, an automated algorithm is presented for comparing the topologies of two circuits.

As follows, two secondary circuits are defined to have *electrically identical topologies (EIT)* if for all the nodes at all the depths of one of the circuits, in the other circuit at the same depth there is a node with equal sets of meters at the node and downstream of the node. This definition allows for different child node ordering between the two trees. The definition is illustrated in Figure 86 that compares an original (true) topology with an electrically identical and non-identical topologies. The middle circuit has the internal order of the node pairs (1,2), (3,4), and (5,6) switched, but has EIT with the original circuit. In the right circuit, nodes 1 and 3 are siblings instead of nodes 1 and 2 being siblings, so these two circuits do not have EIT.

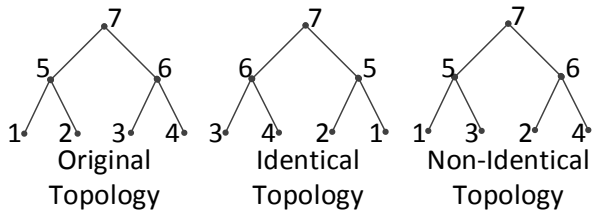

Figure 86. Example of identical and non-identical secondary circuit topologies

Two circuit topologies cannot be directly compared utilizing node-edge incidence matrices because the internal nodes of the estimated circuit tree are unlabeled, i.e., there is no straightforward correspondence between the internal nodes of the two trees. Mathematically, two graphs are defined to be completely structural equivalent if the graphs are related by *isomorphism*, i.e., if there is a structure-preserving vertex bijection between the two graphs [113]. The EIT definition and the definition of isomorphism (of rooted trees) are similar but not identical because the rooted tree isomorphism definition allows graphs with arbitrarily permuted leaf nodes to be related through isomorphism. Figure 87 shows an example of the difference between the EIT and rooted tree isomorphism definitions. The

leaf vertices 1,2,3, and 4 are labeled (metered secondary circuit nodes), whereas the internal vertices are unlabeled (unmetered secondary circuit nodes). A vertex bijection for a possible isomorphism is given in Table 12. The two definitions differ since graph isomorphism does not consider the pre-labeled (metered) leaf nodes.

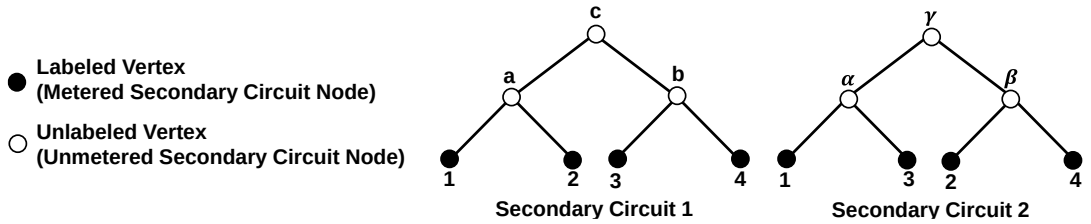

Figure 87. An example of the difference between electrically identical topologies and and graph isomorphism

Table 12. Vertex (node) bijection for a possible isomorphism between secondary circuit trees 1 and 2 in Figure 86

Secondary Circuit 1	Secondary Circuit 2
1	1
2	3
3	3
4	4
a	α
b	β
c	γ

Isomorphism of two rooted trees can be verified with the AHU (Aho, Hopcroft, and Ullman) algorithm, which has a linear complexity with respect to the number of nodes N, i.e., $\mathcal{O}(N)$ [114]. In order to correctly verify if two trees have EITs, the AHU algorithm was modified to consider the leaf node labeling. The resulting pseudocode is listed in Algorithm 6. The algorithm correctly identified the tested various secondary circuit pairs, part of which have EITs and others do not have.

Algorithm 6. Check if two trees have electrically identical topologies

Input: Rooted trees T1 and T2
Output: TRUE or FALSE (do the trees have EIT)
1. IF the trees do not have the same number of nodes N
Return FALSE and **STOP**.
ENDIF
2. For all nodes of T1 and T2, determine L_i, node i distance from the tree root node. Find the maximum node distance from the root node $\overline{L} = \max_{i \in \{1,\dots,N\}} L_i$ and calculate the node depth: $D_i = \overline{L} - L_i$.
3. IF maximum node distances $\overline{L}$ of T1 and T2 are not equal
Return FALSE and **STOP**.
ENDIF
4. Label nodes that have meters with the meter name and nodes without meters with an empty label.
5. FOR all nodes of T1 and T2 at each depth $i = 1, \dots, \overline{L}$
a) Assign each node a label consisting of the sorted child node labels, e.g., if child node labels of a node are "1", "2", "7", then the node is labeled: "((1),(2),(7))".
b) IF the list of the sorted node labels at depth i are equal for T1 and T2 return FALSE and **STOP**.
ENDIF
ENDFOR
6. Return TRUE and **STOP**.

6.5 Topology Estimation Results

6.5.1 Results for the 66-Node Test Circuit

The LRTE algorithm was first analyzed on the three-phase 66-node test circuit introduced in section 4.5.1. The circuit topology is shown in Figure 88.

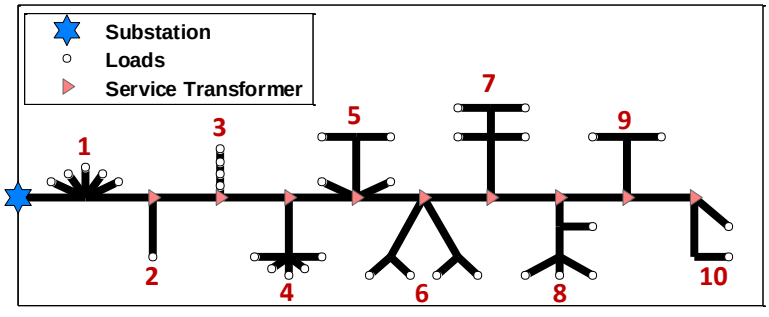

Figure 88. Three-phase test circuit topology (secondary circuit numbers in red)

The LRTE algorithm was utilized to estimate the 10 secondary circuit topologies and parameters. The true and the estimated topologies and component parameters of secondary circuits 5 and 8 are compared in Figure 89. The true and the estimated topologies match perfectly and the estimated impedances are very close to the true impedances.

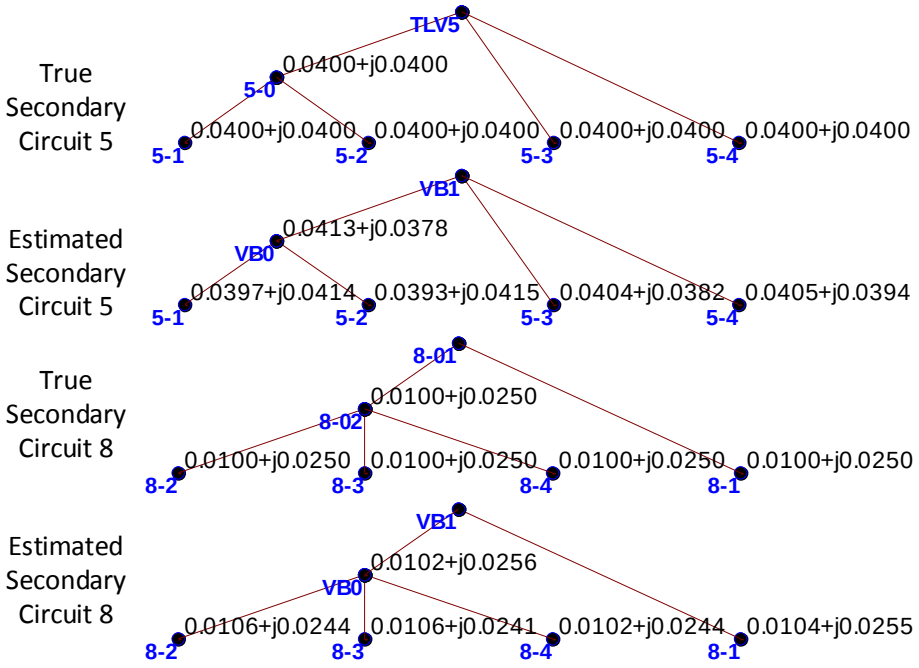

Figure 89. True and estimated topologies of secondary circuits 5 and 8 (node names in bold blue, and node upstream branch impedances in black)

Table 13 lists the average and maximum errors of the *R* and *X* parameters estimated with 8760 measurement samples both without and with measurement error. In both cases, the estimated topologies match perfectly with the true topologies for all the 10 secondary circuits. Without measurement error, most parameters are accurately estimated and even the worst-case accuracy is well within acceptable level. With the practical level: 1% P, 1% Q, and 0.2% V measurement error, most parameters are still estimated with a good accuracy and even the worst-case parameter estimation error of around 18% is an enormous improvement from having no information of the secondary system topology and parameters. The estimated parameters without and with measurement error are visualized

in Figure 90 and Figure 91, respectively. The parameters are estimated with a slightly poorer accuracy compared to the case with known topology that is shown in section 4.6. The slightly poorer accuracy in the unknown topology case is caused mainly by the linear regression parallel branch parameter estimation that is done sequentially with a simpler model (71) in the unknown topology case and is done with a more accurate single model (30)-(33) in the known topology case. In principle, after the secondary circuit topology is estimated with the approach shown in this chapter, the secondary circuit parameters could be re-estimated with the LRPE approach shown in Chapter 4. However, the accuracy improvement may be minor compared to inconsistencies caused by measurement error and model inconsistencies.

Table 13. The average relative errors of the estimated R and X

Meas. Error?	$R_{err,avg}$ [%]	$X_{err,avg}$ [%]	$R_{err,max}$ [%]	$X_{err,max}$ [%]
No	0.45	0.36	2.80	1.44
Yes	3.31	3.84	17.57	11.66

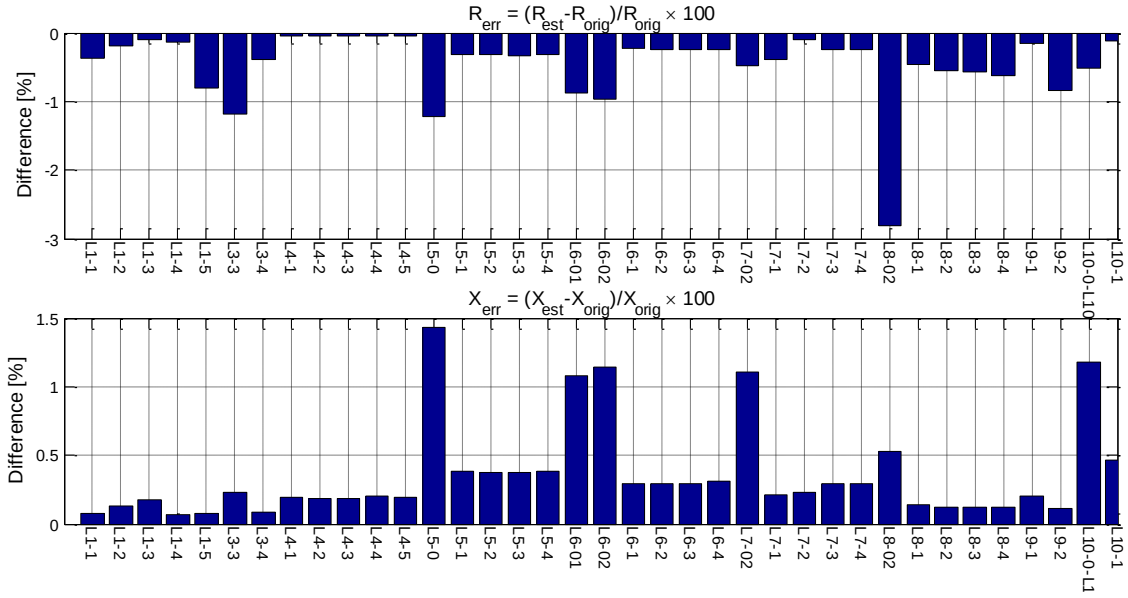

Figure 90. Relative error of estimated R and X without measurement error

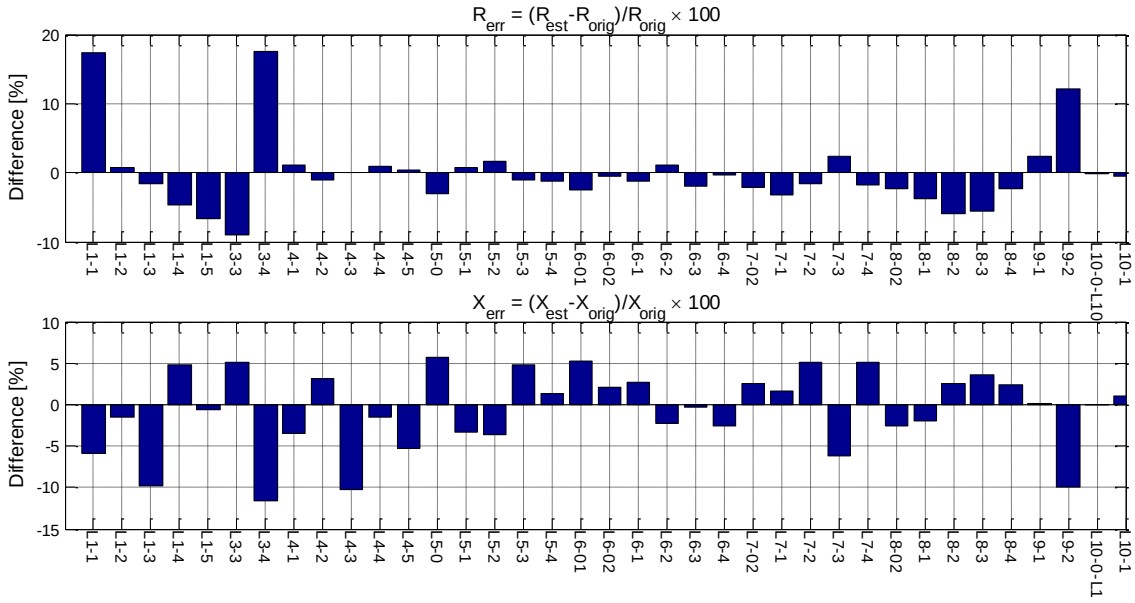

Figure 91. Relative error of estimated R and X with 1% P, 1% Q, and 0.2% V measurement error

6.5.2 Results for a Georgia Tech Feeder

The LRTE algorithm was utilized to estimate the secondary circuit topologies of one of the feeders of the Georgia Tech distribution system presented in section 3.1. Figure 92 shows two examples of the original secondary circuit model and the estimated secondary circuit. For the Georgia Tech system, the secondary system topologies are known from building diagrams, but the impedances, cable types, and lengths are not well known. The LRTE algorithm is not perfect, and in some cases when there is significant measurement error, the topologies are not correctly estimated. Like any other statistical estimation methods, the topology estimation results are strongly data-driven and thus, good quality data is an imperative to receive good results.

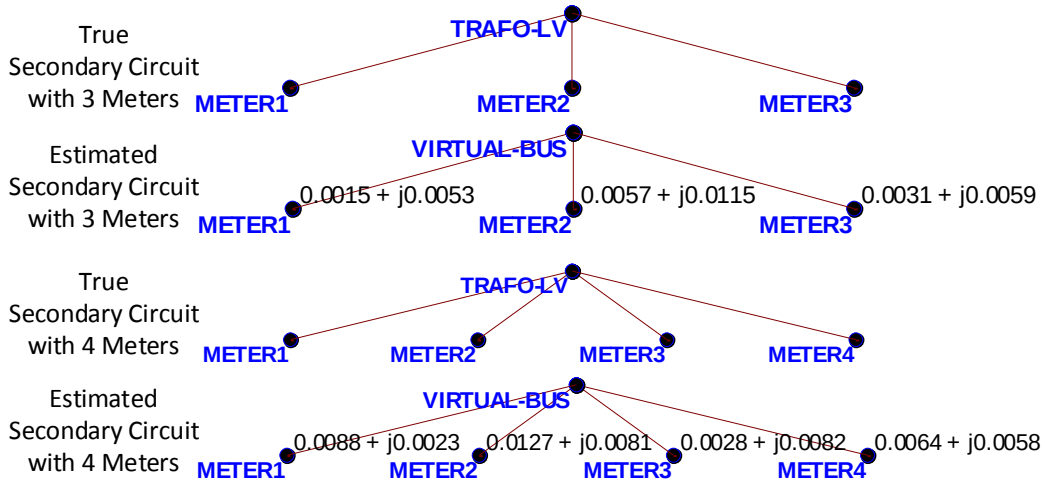

Figure 92. Original and estimated topology of a secondary circuits with three meters and a secondary circuits with four meters

6.6 Discussion

The LRTE algorithm presented in this chapter accurately estimates the test circuit parameters both without and with measurement error. The algorithm also correctly estimates the topologies of most of the Georgia Tech secondary circuits. The most challenging part of LRTE algorithm in Algorithm 5 is step 4, i.e., correctly selecting the linear regression model that provides the best fit. This led to various challenges with high measurement errors. In certain topologies, if the incorrect meter pair was selected even once, the final estimated topology would be wrong. Thus, it is crucial to pair correct meters at each iteration of the LRTE algorithm. Several regression model selection criteria were analyzed including R-squared and root mean squared error (RMSE). R-squared, which is a metric measuring to which degree a regression model describes the variation of the data, seems to prefer selecting meter pairs with larger voltage drop (larger response variable of the regression model) and thus, lower level of relative measurement noise. This resulted in an incorrect topology in many cases. RMSE turned out to be the best model selection criteria resulting in correct estimation of all the secondary circuit topologies in the 66-node test circuit.

6.7 Towards Distribution System Parameter and Topology Estimation

Distribution system secondary circuit parameter and topology estimation have important roles of validating and refining the existing utility feeder models and thus, preparing them for increased situational awareness and operational tasks in the future smart distribution systems. Next, practical implementation of parameter and topology estimation is discussed.

6.7.1 Distribution System Parameter and Topology Estimation Process

The high-level secondary circuit parameter and topology estimation algorithm is shown in Figure 93. The first step is to validate the measurement data and impute the bad and missing measurement samples for the time series power flow analysis with the approaches discussed in Chapter 3. Then, the existing utility feeder model is compiled and time series power flow is solved utilizing load active and reactive power (or current and power factor) measurements, substation voltage measurements (if available), and PV generation as inputs. In this dissertation, the distribution system power flow is solved with OpenDSS, and all parameter and topology estimation algorithms are implemented in MATLAB [115], [116]. The output from the time series power flows solutions are the service transformer MV-side voltages, which are needed to estimate the service transformer impedance. The time series power flow solutions also provide the secondary circuit sensor voltages simulated with original secondary circuit models. In the feeder medium-voltage model validation, the simulated sensor voltages are compared to the measured sensor voltages validating for medium-voltage model gross errors including the voltage regulating device operation modeling, etc. Some of these validation principles are discussed in section 5.2.4.

Next, the algorithm proceeds one secondary circuit at a time, estimating the secondary circuit topologies and branch impedances with the approaches shown in

Chapters 4-6. After all the secondary circuits have been processed, another time series power flow simulation is run with the estimated secondary circuit models to compare measured voltages to the voltages simulated with the estimated secondary circuit models. In the manual verification step, the user needs to compare the estimated parameter values and how closely they align with physically expected values. The manual verification of the estimation results is very important in order to avoid any possibilities of replacing previously accurate impedance parameters with poorer estimates. This step is also useful for detecting any data or topology problems based on, e.g., physically impossible parameter estimates or poor linear regression fits.

It should be emphasized that the parameter and topology estimation methods that are presented in this dissertation do not require modifying any existing utility software. Moreover, the presented methods are computationally highly efficient since no iterative power flow solutions are required during the parameter and topology estimation. Instead, the linear regression parameter and topology estimation only require solving a linear system, which can be done in a fraction of a second even when thousands of measurement samples are leveraged to counteract the accuracy, granularity, and time-synchronization issues related to AMI and DER measurements. Moreover, the presented methods allow processing each secondary circuit individually thus, making it possible to divide large and complicated feeder models to smaller sub problems. This divide and conquer approach significantly reduces the amount of input and output data that needs to be handled simultaneously and thus, allows utilizing large sample sizes for the estimation. Since typical distribution feeder models consist of thousands of lines, hundreds of distribution transformers, and thousands of customers, it is very attractive to perform the estimation for one secondary circuit at a time. Moreover, since typical distribution system operators have hundreds or thousands of distribution feeders, the computational time per feeder model

must remain modest in order for distribution system parameter and topology estimation to be a practical and cost-effective approach for model calibration.

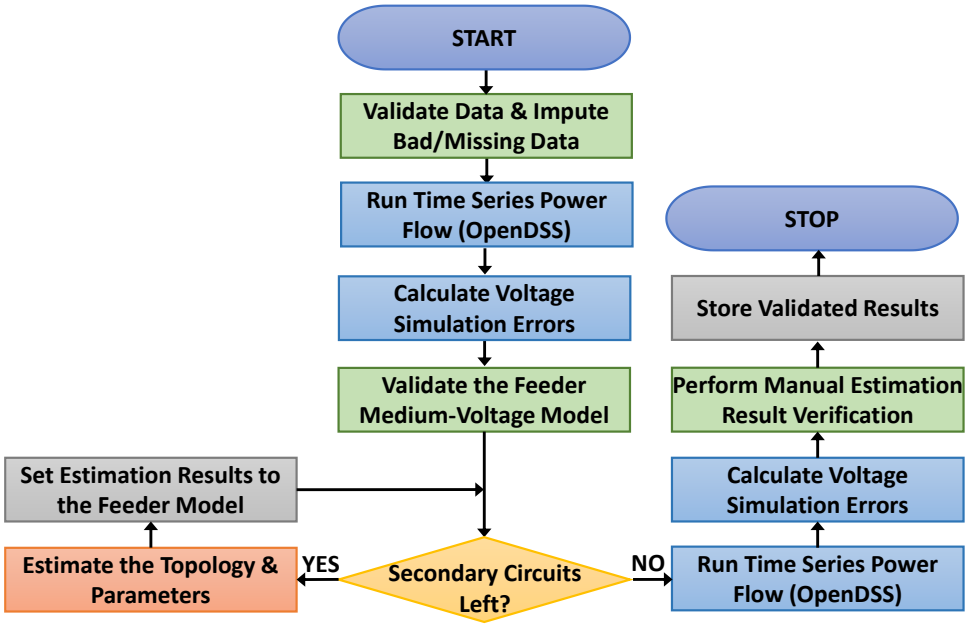

Figure 93. Distribution system parameter and topology estimation process

6.7.2 Leveraging the Big Data for Parameter and Topology Estimation

Next, practical issues for leveraging the Big Data from smart meters and other sensors for distribution system parameter and topology estimation are discussed. Figure 94 illustrates the flows of Big Data for distribution system parameter and topology estimation. The current model components, parameters and permanent connectivity will be fetched from GIS to build the distribution system model. SCADA will transmit the historical device measurements and states. AMI will provide the load profiles, and DER the generation profiles, as an input to time series power flows that simulate the service transformer primary voltages. By leveraging the simulated service transformer voltages and distributed voltage and power (or current) measurements from the AMI and DER as well, the parameter and topology estimator will estimate the secondary system component connectivity and parameters. The estimation results are passed back to the distribution

system model to simulate time series power flows with the estimated parameters and topology. After passing a manual validation, the estimated component parameters and topology are passed to GIS and the distribution system model.

The Big Data challenge is efficiently managing the data flows through advanced data analytics, optimized database queries, and rapid time series analysis. For distribution system parameter and topology estimation to be practical, data processing and analyses must be automated as much as possible with limited human intervention to perform the manual validation of results. Moreover, to allow rapid manual validation of results, primary circuit gross modeling errors, suspected bad parameter and topology estimates, and bad measurement data must be automatically identified.

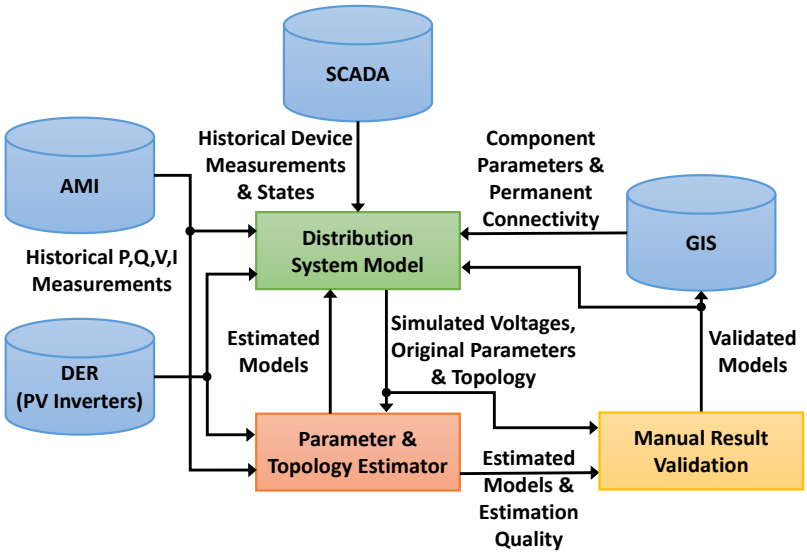

Figure 94. Big Data for distribution system parameter and topology estimation

As parameter and topology estimation is performed offline, measurement system delays are not an issue but poorly synchronized measurement data must be re-synchronized, e.g., using linear (or other) interpolation. Any inaccuracy resulting from the measurement re-synchronization can be counteracted by utilizing historical Big Data with

large sample sizes, which the proposed parameter and topology estimation methods can effectively handle.

Measurement data must also be preprocessed to identify and clean bad and missing data. A meter can have gross errors in some or in all of its measurement samples. Since the proposed estimation methods can easily utilize thousands of samples, some measurement samples with gross errors do not have a high influence on the estimation results. Some bad measurement samples of a meter can be identified with typical methods for detecting outliers in the linear regression response variable or the predictor variables [104]. On the other hand, many cases when all measurements of a meter have gross errors can be easily identified with simple checks such as the ones discussed in section 3.2. Moreover, if all loads in a secondary circuit have smart meters, it may be possible to identify meter (or model) gross errors from poor parameter estimation linear regression model fit (low R-squared values, high RMSE, insignificant parameter p-values). However as discussed in section 4.6, these metrics are not always effective at distinguishing between good and bad regression models. In some cases it can be very hard to identify meters with gross errors, e.g., when a load, which is small compared to the other loads in a secondary circuit, has a meter with gross errors.

Once missing and bad data samples have been identified, they must be imputed to allow the two time-series power flows that require full set of measurement data. Since bad or missing data always results in lost information, the samples during which any secondary circuit meter has missing or bad data should not be used for the actual linear regression parameter and topology estimation but only for running the time-series power flows.

CHAPTER 7. CONCLUSIONS AND CONTRIBUTIONS

7.1 Conclusions

Efficient management of distributed energy resources with advanced Volt/VAr optimization and other advanced distribution management system functions requires accurate and reliable distribution system modeling, monitoring, and coordination. This can be realized by exploiting the Big Data from modern distribution system measurement sources such as AMI and PV micro-inverters.

While modern distribution measurement sources provide more data, they typically are subject to longer delays and have lower measurement granularity, accuracy, and reliability than transmission system SCADA. Accurate and robust use of the modern distribution system measurements will be a cornerstone of the future advanced distribution management systems.

Because of the large number of parameters and load conditions involved in a distribution system model, there is a large degree of uncertainty with respect to the accuracy and quality of current utility models. Stored circuit model including the model parameter values and topology may be incorrect as a result of unknown data, human errors, inaccurate manufacturing data, network changes, incorrect tap information, etc. Improving the accuracy and detail of the feeder models is important to allow ADMS functions with high penetrations of DERs. It is particularly important to improve the models of the distribution system secondary circuits where a large share of the DERs are located.

This dissertation presents intelligent methods for managing the issues related to the modern distribution system measurements and for leveraging the data for distribution model calibration. The presented methods are proven to be efficient using the Georgia Tech distribution system with SCADA and AMI measurements, and large utility feeder models

with SCADA and PV micro-inverter measurements. The methods can be utilized to any distribution system, independent of its size or type.

The developed data validation methods were effectively used to detect numerous issues in Georgia Tech AMI. Similar data issues are expected in other AMIs of this scale. Compared to conventional linear interpolation, historic average, and an industry best practice imputation approaches, the developed optimally weighted average data imputation method had a superior average accuracy in imputing Georgia Tech AMI measurements. The method creates a series of imputed samples that have a continuous profile with respect to the adjacent available measurements, which is a highly desirable feature for time-series analyses. The weight parameter of the developed imputation method is trained offline after which the method is computationally and data efficient making the method suitable for both offline and online settings.

Four distribution secondary circuit parameter and topology estimation algorithms were developed to handle known and unknown secondary circuit topologies with different sets of available measurements. Table 14 summarizes the algorithms and the required measurements.

Table 14. The developed distribution secondary circuit estimation algorithms

Algorithm	Requires Known Topology	Requires Power Measurements	Requires Voltage Measurements
LRPE (Chapter 4)	Yes	Yes	Yes
LRPE with meters not reporting voltage measurements (Section 5.1)	Yes	Yes	At least one sensor
SLRPE (Section 5.2)	No	No	At least one sensor
LRTE (Chapter 6)	No	Yes	Yes

The developed algorithms offer accurate, flexible, and computationally efficient approaches for enhancing distribution model detail and accuracy. The algorithms can be run separately for each secondary circuit and can be solved in seconds despite utilizing thousands of measurement samples to counteract the accuracy and granularity issues

related to the modern measurements. Moreover, it should be emphasized that the developed parameter and topology estimation methods do not require modifying any existing utility software.

The developed LRPE algorithm estimates three-phase test circuit secondary parameters without (with) practical levels of measurement error with an average error less than 1% (3%). The method also effectively reduces the mean bias errors from the Georgia Tech simulated load voltages. The secondary circuit parameter estimation requires accurate voltage measurements but the influence of measurement error can be reduced by using large sample sizes, which the proposed method can easily handle.

The LRPE algorithm with meters not reporting voltage measurements expands the LRPE algorithm applicability to cases when some meters do not provide voltage measurements. The algorithm accuracy is comparable to the fully available AMI case although the accuracy depends somewhat on the secondary circuit topology and the number of meters reporting voltages. The algorithm performs better if the meters that do not report voltages, report current and power factor measurements.

The developed SLRPE algorithm estimates the simplified secondary circuit parameters with an average error less than 0.4% provided that accurate power and voltage measurements are available. When loads are modeled through inaccurate substation load allocation, the method reduces the average absolute PV voltage simulation errors for three tested utility feeder models on average by 19.3%, 22.5%, and 71.5%. However, the results indicate that load allocation should not be used for (secondary circuit) parameter estimation. The results also underline the importance of accurate feeder medium-voltage modeling.

The developed LRTE algorithm accurately estimates the three-phase test case secondary circuit parameters both without and with measurement error. The algorithm also correctly estimates the topologies of most of the Georgia Tech secondary circuits. The most

challenging part of the LRTE algorithm is to correctly select the linear regression model that provides the best fit and leads to the correct topology, which is especially challenging under conditions with high measurement errors.

Many utilities have extensively rolled out AMI systems, mostly for automatic billing purposes. This dissertation presents several new use cases for the AMI and DER sensor data and thus, encourages further investments in AMI and other modern distribution sensors. Additionally, the presented work addresses the need for utilities to manage the Big Data from modern measurement sources and to utilize it to improve the analytical and operational modeling accuracy for future smart distribution systems with advanced distribution management system functions and ubiquitous distributed energy resources.

7.2 Contributions

The work presented in this dissertation has led to several first author publications, including two journal publications ([91], [98]), four conference publications ([42], [92], [107], [111]), and a technical report ([97]). The work has also contributed to another technical report ([99]). Furthermore, during his PhD program Mr. Peppanen has also authored or co-authored four other conference papers ([12], [117]–[119]).

In summary, the primary contributions of the presented work are:

- Demonstration of smart meter data validation methods on the Georgia Tech distribution system AMI.
- Detection of numerous issues in Georgia Tech AMI measurements.
- Implementation of Georgia Tech distribution system model and validation of the model with AMI data.
- Implementation of a Matlab toolbox for easy access, validation, and imputation of the Georgia Tech AMI measurements.

- Development of a computationally and data efficient method for imputing missing, bad, and delayed measurements for off-line and online purposes. Demonstration of the algorithm on the Georgia Tech AMI data.
- Development of an accurate, flexible, and computationally efficient distribution secondary circuit parameter estimation method with fully available AMI or other sensor measurements.
- Expansion of the parameter estimation method to handle sensors that do not report voltage measurements.
- Development of a flexible and computationally efficient method for generating secondary circuit models when limited measurements are available in the secondary circuits.
- Generation of 1-phase secondary circuit models with PV micro-inverter measurements for three large U.S. utility feeder models.
- Development of secondary circuit joint topology and parameter estimation algorithm with fully available AMI or other sensor measurements and demonstration of the algorithm on one of the Georgia Tech feeders.

7.3 Recommended Future Work

There are several interesting directions for future work on data imputation. First, the developed data imputation method could be further improved to better handle short missing data periods. This could be achieved by, e.g., utilizing an offset parameter that forces all the weight for the linear interpolation imputation for very short missing data intervals. The relationship between the proposed imputation method and conventional statistical prediction model averaging methods such as stacking could also be investigated. Moreover, the proposed imputation method could be extended to leverage the missing data period energy readings (when available).

The parameter and topology estimation algorithms should be further developed to handle some unmetered secondary circuit loads and/or DERs that report no voltage or power measurements. The parameter and topology estimation algorithm should also be developed to automatically detect meters that are placed in wrong secondary circuits in the GIS.

This dissertation addresses some of the feeder medium-voltage modeling challenges but more work is needed to fully leverage the modern distribution measurements for distribution medium-voltage model calibration. The important future work areas include estimation of lateral phasing, service transformer connections, and the historical states of feeder substation LTC, voltage regulators, and capacitor banks.

The linear regression parameter and topology estimation methods presented in this dissertation are based on a linearized voltage drop approximation. Although the approximation is shown to be accurate for most operating conditions, it requires using the hierarchical parameter and topology estimation approaches shown in this dissertation. Future work should study the accuracy of the hierarchical approach for medium-voltage circuit topology and parameter estimation. Alternative linear and nonlinear formulations should also be considered.

All the developed parameter and topology estimation should be analyzed on more feeders that have AMI, PV micro-inverter, and other sensor measurements.

REFERENCES

[1] "Technology Roadmap Solar Photovoltaic Energy 2014 Edition," International Energy Agency, 2014.

[2] "SEIA Solar Energy Industries Association," 2013. [Online]. Available: http://www.seia.org/. [Accessed: 10-Dec-2013].

[3] "Annual Energy Outlook 2015 with projections to 2040," U.S. Energy Information Administration, Apr. 2015.

[4] "California Solar Statistics." [Online]. Available: https://www.californiasolarstatistics.ca.gov/. [Accessed: 29-Apr-2015].

[5] "California Solar," *SEIA*. [Online]. Available: http://www.seia.org/state-solar-policy/california.

[6] Y. Liu, J. Bebic, B. Kroposki, J. de Bedout, and W. Ren, "Distribution System Voltage Performance Analysis for High-Penetration PV," in *IEEE Energy 2030 Conference*, 2008, pp. 1–8.

[7] J. Quiroz and M. Reno, "Detailed grid integration analysis of distributed PV," in *2012 38th IEEE Photovoltaic Specialists Conference (PVSC)*, 2012, pp. 596–601.

[8] A. Ipakchi and F. Albuyeh, "Grid of the future," *IEEE Power Energy Mag.*, vol. 7, no. 2, pp. 52–62, Mar. 2009.

[9] R. F. Arritt and R. C. Dugan, "Distribution System Analysis and the Future Smart Grid," *IEEE Trans. Ind. Appl.*, vol. 47, no. 6, 2011.

[10] M. E. Baran, H. Hooshyar, Z. Shen, and A. Huang, "Accommodating High PV Penetration on Distribution Feeders," *IEEE Trans. Smart Grid*, vol. 3, no. 2, pp. 1039–1046, Jun. 2012.

[11] "Oahu Locational Value Map," *Hawaiian Electric Maui Electric Hawai'i Electric Light*. [Online]. Available: https://www.hawaiianelectric.com/clean-energy-hawaii/integration-tools-and-resources/locational-value-maps/oahu-locational-value-map-(lvm). [Accessed: 27-Apr-2015].

[12] J. Peppanen and S. Grijalva, "Neighborhood electric vehicle charging scheduling using particle swarm optimization," in *IEEE PES General Meeting*, 2014, pp. 1–5.

[13] A. Ipakchi and F. Albuyeh, "Grid of the future," *IEEE Power Energy Mag.*, vol. 7, no. 2, 2009.

[14] R. Liu, L. Dow, and E. Liu, “A survey of PEV impacts on electric utilities,” in *Innovative Smart Grid Technologies (ISGT), 2011 IEEE PES*, 2011, pp. 1–8.

[15] J. Taylor, A. Maitra, M. Alexander, D. Brooks, and M. Duvall, “Evaluations of plug-in electric vehicle distribution system impacts,” in *IEEE Power and Energy Society General Meeting*, 2010.

[16] Q. Gong, S. Midlam-Mohler, V. Marano, and G. Rizzoni, “Study of PEV Charging on Residential Distribution Transformer Life,” *IEEE Trans. Smart Grid*, vol. 3, no. 1, pp. 404–412, Mar. 2012.

[17] L. Pieltain Fernández, T. G. S. Román, R. Cossent, C. M. Domingo, and P. Frías, “Assessment of the Impact of Plug-in Electric Vehicles on Distribution Networks,” *IEEE Trans. Power Syst.*, vol. 26, no. 1, pp. 206–213, Feb. 2011.

[18] K. Clement-Nyns, E. Haesen, and J. Driesen, “The Impact of Charging Plug-In Hybrid Electric Vehicles on a Residential Distribution Grid,” *IEEE Trans. Power Syst.*, vol. 25, no. 1, pp. 371–380, 2010.

[19] “Global EV Outlook - Understanding the Electric Vehicle Landscape to 2020,” International Energy Agency, Apr. 2013.

[20] Y.-F. Huang, S. Werner, J. Huang, N. Kashyap, and V. Gupta, “State estimation in electric power grids: meeting new challenges presented by the requirements of the future grid,” *IEEE Signal Process. Mag.*, vol. 29, no. 5, pp. 33–43, Sep. 2012.

[21] M. Farivar, C. R. Clarke, S. H. Low, and K. M. Chandy, “Inverter VAR control for distribution systems with renewables,” in *2011 IEEE International Conference on Smart Grid Communications (SmartGridComm)*, 2011, pp. 457–462.

[22] J. Fan and S. Borlase, “The evolution of distribution,” *IEEE Power Energy Mag.*, vol. 7, no. 2, pp. 63–68, Mar. 2009.

[23] J. S. John, “Can Microinverters Stabilize Hawaii’s Shaky Grid? : Greentech Media,” 02-Feb-2015. [Online]. Available: http://www.greentechmedia.com/articles/read/enphase-to-help-hawaii-ride-its-solar-energy-wave. [Accessed: 03-Feb-2015].

[24] R. F. Arritt, R. C. Dugan, R. W. Uluski, and T. F. Weaver, “Investigation load estimation methods with the use of AMI metering for distribution system analysis,” in *2012 IEEE Rural Electric Power Conference (REPC)*, 2012, pp. B3–1–B3–9.

[25] W. H. Kersting and W. H. Phillips, “Load allocation based upon automatic meter readings,” presented at the Transmission and Distribution Conference and Exposition, 2008, pp. 1–7.

[26] "VOICES of Experience - Insights into Advanced Distribution Management Systems," Feb. 2015.

[27] X. Feng, F. Yang, and W. Peterson, "A practical multi-phase distribution state estimation solution incorporating smart meter and sensor data," in *IEEE Power and Energy Society General Meeting*, San Diego, CA, 2012.

[28] T. Taylor and H. Kazemzadeh, "Integrated SCADA/DMS/OMS: Increasing Distribution Operations Efficiency," *Electric Energy Online*, Apr-2009.

[29] M. Kezunovic, "Smart Fault Location for Smart Grids," *IEEE Trans. Smart Grid*, vol. 2, no. 1, pp. 11–22, Mar. 2011.

[30] "Report on Use of Distribution State Estimation Results for Distribution Network Automation Functions, D2.2.1," BU, EF, Fraunhofer IWES, INDRA, Korona, FP7 - 248135, Apr. 2011.

[31] D. A. Haughton and G. T. Heydt, "A linear state estimation formulation for smart distribution systems," *IEEE Trans. Power Syst.*, vol. 28, no. 2, pp. 1187–1195, May 2013.

[32] J. Medina, N. Muller, and I. Roytelman, "Demand Response and Distribution Grid Operations: Opportunities and Challenges," *IEEE Trans. Smart Grid*, vol. 1, no. 2, pp. 193–198, 2010.

[33] "Smart meters and smart meter systems: a metering industry perspective," Edison Electric Institute (EEI), Association of Edison Illuminating Companies (AEIC), Utilities Telecom Council (UTC), Washington DC, EEI-AEIC-UTC White Paper, Mar. 2011.

[34] M. Harnish, "Electricity Monthly Update - February 2015," Energy Information Agency, Apr. 2015.

[35] "ANSI 12.10-2011 Physical Aspects of Watthour Meters - Safety Standard." American National Standard Institute (ANSI).

[36] "ANSI 12.20-2010 for Electricity Meters - 0.2 and 0.5 Accuracy Classes." American National Standard Institute (ANSI).

[37] "ANSI C12.1-2008 Electric Meters Code for Electricity Metering." American National Standards Institute (ANSI), 27-Jun-2008.

[38] "IEEE Std C57.13-2008 - IEEE Standard Requirements for Instrument Transformers," *IEEE Std C5713-2008 Revis. IEEE Std C5713-1993*, pp. c1–82, Jul. 2008.

[39] "Accuracy of Digital Electricity Meters," Electric Power Research Institute, May 2010.

[40] L. A. Irwin, "A high accuracy standard for electricity meters," Schneider Electric, White paper, Apr. 2011.

[41] F. P. Sioshansi, *Smart grid integrating renewable, distributed & efficient energy*. Waltham, MA: Academic Press, 2012.

[42] J. Peppanen, J. Grimaldo, M. J. Reno, S. Grijalva, and R. G. Harley, "Increasing distribution system model accuracy with extensive deployment of smart meters," in *IEEE Power and Energy Society General Meeting*, Washington D.C., 2014.

[43] "Demand Reductions from the Application of Advanced Metering Infrastructure, Pricing Programs, and Customer-Based Sstems - Initial Results," Department of Energy, Smart Grid Investment Grant Program Report, Dec. 2012.

[44] "An Assessment of Interval Data and Their Potential Application to Residential Electricity End-Use Modeling," Energy Information Agency, Washington D.C., Feb. 2015.

[45] A. P. S. Meliopoulos, G. J. Cokkinides, R. Huang, E. Farantatos, S. Choi, Y. Lee, and X. Yu, "Smart Grid Technologies for Autonomous Operation and Control," *IEEE Trans. Smart Grid*, vol. 2, no. 1, pp. 1–10, 2011.

[46] A. Abur, *Power system state estimation: theory and implementation*. New York, NY: Marcel Dekker, 2004.

[47] W. Luan, J. Peng, M. Maras, and J. Lo, "Distribution network topology error correction using smart meter data analytics," in *IEEE Power and Energy Society General Meeting*, Vancouver, BC, Canada, 2013.

[48] J. A. Taylor, T. A. Short, and B. Bushey, "Efficiency impacts of distribution secondaries," in *IEEE Transmission and Distribution Conference and Exposition*, Orlando, FL, USA, 2012, pp. 1–6.

[49] G. J. Shirek, B. A. Lassiter, W. C. Carr, and W. H. Kersting, "Modeling Secondary Services in Engineering and Mapping," *IEEE Trans. Ind. Appl.*, vol. 48, no. 1, pp. 254–262, Jan. 2012.

[50] "Uniform Business Practices for Unbundled Electricity Metering - Volume 2," Edision Electric Institute, Dec. 2000.

[51] J. Chen, W. Li, A. Lau, J. Cao, and K. Wang, "Automated Load Curve Data Cleansing in Power Systems," *IEEE Trans. Smart Grid*, vol. 1, no. 2, pp. 213–221, Sep. 2010.

[52] P. Allison, “Missing Data,” in *The SAGE Handbook of Quantitative Methods in Psychology*, Thousand Oaks, CA: Sage Publications Inc., 2009, p. 18.

[53] X.-H. Zhou, *Applied missing data analysis in the health sciences*. Hoboken, New Jersey: John Wiley & Sons, Inc, 2014.

[54] Little, Roderick and D. B. Rubin, *Statistical Analysis with Missing Data*, 2nd ed. 2002.

[55] H. Junninen, H. Niska, K. Tuppurainen, J. Ruuskanen, and M. Kolehmainen, “Methods for imputation of missing values in air quality data sets,” *Atmos. Environ.*, vol. 38, no. 18, Jun. 2004.

[56] I. Žliobaitė and J. Hollmén, “Optimizing regression models for data streams with missing values,” *Mach. Learn.*, pp. 1–27, Jul. 2014.

[57] X. Lu, J. Si, L. Pan, and Y. Zhao, “Imputation of missing data using ensemble algorithms,” in *2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, 2011, vol. 2, pp. 1312–1315.

[58] H. Alfares and M. Nazeerudding, “Electric load forecasting: literature survey and classification of methods,” *Int. J. Syst. Sci.*, vol. 33, no. 1, pp. 23–34, 2002.

[59] T. Hong, “Short Term Electric Load Forecasting,” PhD Dissertation, North Carolina State University, Raleigh, NC, 2010.

[60] G. Gross and F. D. Galiana, “Short-term load forecasting,” *Proc. IEEE*, vol. 75, no. 12, pp. 1558–1573, Dec. 1987.

[61] R. A. Sevlian and R. Rajagopal, “A Model For The Effect of Aggregation on Short Term Load Forecasting,” presented at the IEEE Power and Energy Society General Meeting, Washington D.C., 2014.

[62] A. Seppala, “Statistical distribution of customer load profiles,” in *, 1995 International Conference on Energy Management and Power Delivery, 1995. Proceedings of EMPD ’95*, 1995, vol. 2, pp. 696–701 vol.2.

[63] R. Li, C. Gu, Y. Zhang, and F. Li, “Implementation of load profile test for electricity distribution networks,” in *2012 IEEE Power and Energy Society General Meeting*, 2012, pp. 1–6.

[64] U. Singh and M. E. Baran, “Load Estimation for distribution feeder monitoring and management,” in *2010 IEEE Power and Energy Society General Meeting*, 2010, pp. 1–6.

[65] B. Stephen, A. J. Mutanen, S. Galloway, G. Burt, and P. Jarventausta, "Enhanced Load Profiling for Residential Network Customers," *IEEE Trans. Power Deliv.*, vol. 29, no. 1, Feb. 2014.

[66] B. Stephen and S. J. Galloway, "Domestic Load Characterization Through Smart Meter Advance Stratification," *IEEE Trans. Smart Grid*, vol. 3, no. 3, pp. 1571–1572, Sep. 2012.

[67] R. Singh, B. C. Pal, and R. A. Jabr, "Distribution system state estimation through Gaussian mixture model of the load as pseudo-measurement," *IET Gener. Transm. Distrib.*, vol. 4, no. 1, pp. 50–59, Jan. 2009.

[68] G. Mateos and G. B. Giannakis, "Load Curve Data Cleansing and Imputation Via Sparsity and Low Rank," *IEEE Trans. Smart Grid*, vol. 4, no. 4, pp. 2347–2355, Dec. 2013.

[69] X. Zhang, S. Grijalva, and M. J. Reno, "A time-variant load model based on smart meter data mining," in *IEEE PES General Meeting*, Washington DC, 2014, pp. 1–5.

[70] A. D. Domınguez-Garcıa, G. T. Heydt, and S. Suryanarayanan, "Implications of the Smart Grid Initiative on Distribution Engineering Final Project Report - Part 3 Restoration, State Estimation and Reliability Enhancement," PSERC Publication 11-06, Sep. 2011.

[71] M. Baran and T. E. McDermott, "State estimation for real time monitoring of distribution feeders," in *IEEE Power and Energy Society General Meeting*, Calgary, AB, Canada, 2009.

[72] M. Baran and T. E. McDermott, "Distribution system state estimation using AMI data," in *IEEE Power Systems Conference and Exposition*, Seattle, WA, 2009.

[73] H. Wang and N. N. Schulz, "A load modeling algorithm for distribution system state estimation," in *IEEE PES Transmission and Distribution Conference and Exposition*, 2001, pp. 102–105.

[74] E. Manitsas, R. Singh, B. Pal, and G. Strbac, "Modelling of pseudo-measurements for distribution system state estimation," in *IET-CIRED SmartGrids for Distribution*, Frankfurt, Germany, 2008.

[75] P. Zarco and A. Gomez-Exposito, "Power system parameter estimation: a survey," *IEEE Trans. Power Syst.*, vol. 15, no. 1, pp. 216–222, Feb. 2000.

[76] A. S. Debs, "Parameter estimation for power systems in the steady-state," in *1974 IEEE Conference on Decision and Control including the 13th Symposium on Adaptive Processes*, 1974, vol. 13, pp. 587–592.

[77] M. R. M. Castillo, J. B. A. London, and N. G. Bretas, “An approach to power system branch parameter estimation,” in *IEEE Electric Power Conference*, Vancouver, BC, Canada, 2008, pp. 1–5.

[78] J. C. S. Souza, A. M. Leite da Silva, and A. P. Alves da Silva, “Online topology determination and bad data suppression in power system operation using artificial neural networks,” *IEEE Trans. Power Syst.*, vol. 13, no. 3, pp. 796–803, Aug. 1998.

[79] Y.-L. Lo, S.-C. Huang, and C.-N. Lu, “Transformational Benefits of AMI Data in Transformer Load Modeling and Management,” *IEEE Trans. Power Deliv.*, vol. Early Access Online, 2013.

[80] A. J. Berrisford, “A tale of two transformers: An algorithm for estimating distribution secondary electric parameters using smart meter data,” in *Annual IEEE Canadian Conference on Electrical and Computer Engineering*, Regina, SK, Canada, 2013.

[81] B. Wang and W. Luan, “Generate distribution circuit model using AMI data,” in *China International Conference on Electricity Distribution*, 2014, pp. 1251–1255.

[82] T. A. Short, “Advanced metering for phase identification, transformer identification, and secondary modeling,” *IEEE Trans. Smart Grid*, vol. 4, no. 2, pp. 651–658, 2013.

[83] M. R. M. Castillo, J. B. A. London, and N. G. Bretas, “Network branch parameter validation based on a decoupled State/Parameter Estimator and historical data,” in *PowerTech, IEEE Bucharest*, 2009.

[84] G. Cavraro, R. Arghandeh, G. Barchi, and A. von Meier, “Distribution network topology detection with time-series measurements,” in *Innovative Smart Grid Technologies Conference, IEEE*, 2015, pp. 1–5.

[85] G. Cavraro, R. Arghandeh, and A. von Meier, “Distribution Network Topology Detection with Time Series Measurement Data Analysis,” *ArXiv150405926 Cs*, Apr. 2015.

[86] A. S. Raffi and R. Rajagopal, “Feeder Topology Identification,” *arXiv:1503.07224v2*, Mar. 2015.

[87] G. Cavraro, R. Arghandeh, K. Poolla, and A. von Meier, “Data-Driven Approach for Distribution Network Topology Detection,” *arXiv:1504.00724v1*, Apr. 2015.

[88] “PQube,” *Power Sensor Ltd.* [Online]. Available: http://www.powersensorsltd.com/. [Accessed: 12-Feb-2016].

[89] W. Luan, J. Peng, M. Maras, J. Lo, and B. Harapnuk, "Smart Meter Data Analytics for Distribution Network Connectivity Verification," *IEEE Trans. Smart Grid*, vol. PP, no. 99, pp. 1–1, 2015.

[90] S. Bolognani, N. Bof, D. Michelotti, R. Muraro, and L. Schenato, "Identification of power distribution network topology via voltage correlation analysis," in *2013 IEEE 52nd Annual Conference on Decision and Control (CDC)*, 2013, pp. 1659–1664.

[91] J. Peppanen, M. J. Reno, M. Thakkar, S. Grijalva, and R. G. Harley, "Leveraging AMI Data for Distribution System Model Calibration and Situational Awareness," *IEEE Trans. Smart Grid*, Jan. 2015.

[92] J. Peppanen, M. J. Reno, X. Zhang, and S. Grijalva, "Handling Bad or Missing Smart Meter Data through Advanced Data Imputation," presented at the Innovative Smart Grid Technologies, Minneapolis, MN, 2016.

[93] W. H. Kersting, *Distribution system modeling and analysis*. Boca Raton, FL: CRC Press, 2002.

[94] "ANSI C84.1-2011 American national standard for electric power systems and equipment - voltage ratings (60 Hertz)." American National Standard Institute (ANSI), Dec-2006.

[95] Broderick, Robert J., J. E. Quiroz, Matthew J. Reno, A. Ellis, J. Smith, and R. C. Dugan, "Time Series Power Flow Analysis for Distribution Connected PV Generation," Sandia National Laboratories, Albuquerque, New Mexico, SAND2013-0537, Jan. 2013.

[96] M. J. Reno, C. W. Hansen, and J. S. Stein, "Global Horizontal Irradiance Clear Sky Models: Implementation and Analysis," *Sandia Natl. Lab.*, no. SAND2012–2389, 2012.

[97] J. Peppanen, M. J. Reno, R. Broderick, and S. Grijalva, "Distribution System Secondary Circuit Parameter Estimation for Model Calibration," Sandia National Laboratories, SAND2015-7477, Sep. 2015.

[98] J. Peppanen, M. J. Reno, R. Broderick, and S. Grijalva, "Distribution System Model Calibration with Big Data from AMI and PV Inverters," *IEEE Trans. Smart Grid*, pp. 1–10, Mar. 2016.

[99] M. J. Reno, J. Peppanen, J. Seuss, M. Lave, R. J. Broderick, and S. Grijalva, "Tools for Enhanced Grid Operation and Optimized PV Penetration Utilizing Highly Distributed Sensor Data," Sandia National Laboratories, SAND2015-10303, Nov. 2015.

[100] T. Gönen, *Electric power distribution system engineering*, 2nd ed. Boca Raton, FL: CRC Press, 2008.

[101] R. P. Fanning, "Implementation of networked primary and secondary distribution systems for US utilities," in *IEEE Power Engineering Society General Meeting, 2003*, 2003, vol. 4, p. 2429 Vol. 4.

[102] W. H. Kersting, "Center tapped transformer and 120/240 volt secondary models," in *2008 IEEE Rural Electric Power Conference*, 2008, pp. A1–A1–7.

[103] C. W. Brice, "Comparison of Approximate and Exact Voltage Drop Calculations for Distribution Lines," *IEEE Trans. Power Appar. Syst.*, vol. PAS-101, no. 11, pp. 4428–4431, 1982.

[104] M. H. Kutner, C. Nachtsheim, J. Neter, and W. Li, *Applied linear statistical models*. Boston: McGraw-Hill Irwin, 2005.

[105] W. Hołubowski, D. Kurzyk, and T. Trawinski, "A Fast Method for Computing the Inverse of Symmetric Block Arrowhead Matrices," *Appl. Math. Inf. Sci.*, vol. 9, no. 2, pp. 319–324, 2015.

[106] "Pecan Street Inc.," *Pecan Street Inc.* [Online]. Available: http://www.pecanstreet.org/about/.

[107] J. Peppanen, M. J. Reno, R. Broderick, and S. Grijalva, "Secondary Circuit Model Generation Using Limited PV Measurements and Parameter Estimation," in *IEEE PES General Meeting*, Boston, MA, USA, 2016, pp. 1–5.

[108] "NEMA Smart Grid Standards Publication SG-AMI 1-2009 - Requirements for Smart Meter Upgradeability," National Electrical Manufacturers Association, 2009.

[109] M. J. Reno, K. Coogan, R. Broderick, and S. Grijalva, "Reduction of distribution feeders for simplified PV impact studies," in *Photovoltaic Specialists Conference*, Tampa Bay, FL, USA, 2013, pp. 2337–2342.

[110] H. L. Willis, *Power distribution planning reference book*, 2nd ed., rev. and expanded. New York: M. Dekker, 2004.

[111] J. Peppanen, M. J. Reno, R. Broderick, and S. Grijalva, "Distribution System Low-Voltage Circuit Topology Estimation using Smart Metering Data," in *IEEE PES T&D Conference & Exposition*, Dallas, TX, USA, 2016, pp. 1–5.

[112] J. Felsenstein, "The Number of Evolutionary Trees," *Syst. Zool.*, vol. 27, no. 1, pp. 27–33, Mar. 1978.

[113] J. A. Bondy and U. S. R. Murty, *Graph Theory with Applications*, 5th ed. New York, NY: Elsevier Science Publishing Co. Inc., 1976.

[114] A. V. Aho, J. E. Hopcroft, and J. D. Ullman, *The design and analysis of computer algorithms*. Reading, Mass: Addison-Wesley Pub. Co, 1974.

[115] M. J. Reno and K. Coogan, "Grid Integrated Distributed PV (GridPV) Version 2," Sandia National Laboratories, SAND2014-20141, 2014.

[116] "Open Distribution System Simulator," *EPRI*. [Online]. Available: http://sourceforge.net/projects/electricdss/.

[117] J. A. Peppanen, T. Alquthami, D. Molina, and R. Harley, "Optimal PMU placement with Binary PSO," in *IEEE Energy Conversion Conference & Exhibition*, 2012.

[118] M. J. Reno, K. Coogan, J. Peppanen, and S. Grijalva, "Using distribution LMP and time-of-delivery pricing to promote optimal placement and increased profitability of residential PV systems," in *North American Power Symposium*, 2014, pp. 1–6.

[119] J. Peppanen, M. J. Reno, and S. Grijalva, "Thermal energy storage for air conditioning as an enabler of residential demand response," in *North American Power Symposium*, Pullman, WA, 2014, pp. 1–6.